Universidade Nova de Lisboa
Faculdade de Ciências e Tecnologia
Departamento de Informática

Classificação de Documentos

Por
Filipa Alexandra Peleja Madureira 28034

Dissertação apresentada na Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa para obtenção de grau de Mestre em Engenharia de Informática

Orientador: Professor Doutor José Gabriel Pereira Lopes

Lisboa
2009

Dedicatória

À minha mãe, Elisabete Peleja.

Ao meu pai, Henrique Madureira.

Às minhas irmãs, Joana e Rute.

Ao meu Pê.

À minha Kishinha.

Agradecimentos

A todos os que fiz menção na dedicatória, pela paciência e apoio enquanto este trabalho foi realizado. Em especial para a minha mãe.

À minha Kisha que sempre me acalmou nos momentos mais difíceis.

Agradecimentos ao Professor Doutor Gabriel Pereira Lopes por toda a ajuda, e incentivo, disponibilizada.

Resumo

No presente trabalho de investigação pretende-se automatizar o processo de classificação temática de documentos. Foram utilizadas três técnicas de selecção de termos, com três classificadores automáticos, e sete representações de documentos: palavra, multi-palavra, pentagrama, e cadeias dos primeiros 4, 5 e 6 caracteres individualmente, e globalmente.

Entre as técnicas de selecção de termos encontra-se a medida do Terceiro Momento em relação à média. Esta medida foi recentemente proposta, por o Professor Joaquim Ferreira da Silva, e considerou-se importante realizar um estudo comparativo da sua performance em relação a outras medidas, já muito conhecidas e comprovada a sua aplicabilidade. As medidas escolhidas foram: *Chi-Square* e *Information Gain*.

Existem medidas de selecção de termos que demonstram melhores resultados conforme o classificador utilizado, e por isso, as medidas foram experimentadas com diferentes classificadores: K-*Nearest Neighbour*, *RIPPER* e *Support Vector Machines*. São classificadores que na área de classificação demonstraram bons resultados, e assim, avaliou-se o seu desempenho com as diferentes medidas de selecção de termos.

Nos resultados experimentais, em que foi utilizado o *corpus* da *Reuters*-21578, pode-se observar que o desempenho obtido com a técnica do terceiro momento é superior, ou equivalente, à obtida com as medidas de selecção de termos *Chi-Square* e *Information Gain*.

Utilizando diferentes representações de documentos é possível obter um desempenho, com os três classificadores, equivalente ao obtido com a representação de documentos por palavra.

Abstract

This work of investigation aims to automate the process of thematic classification of documents. Three techniques of features selection will be used, with three classifiers automatic and seven representations of documents: word; multi-word; pentagram; and a chain of the first 4, 5 and 6 characters individually and globally.

Among the techniques of features selection there is the Third Moment in relation to the average. This measure has been recently proposed by Professor Joaquim Ferreira da Silva and it is important to conduct a comparative study of its performance in relation to other measures already very well known, whose applicability has also been attested. The chosen measures were *Chi-Square* and *Information Gain*.

There are measures of features selection that demonstrate better results according to the classifier used and therefore measures with different classifiers will be studied: K-*Nearest Neighbour*, *RIPPER* and *Support Vector Machines*. These are classifiers which in the field of classification demonstrate good results having in mind the evaluation of their performance towards the different measures of features selection.

The Corpus Reuters-21578 was used in the experimental results allowing us to observe that the performance obtained by the third moment technique is higher or equivalent to the one obtained by the others measures of terms selection.

Using different representations of documents it is possible to obtain a performance with the three classifiers equivalent to the one obtained by the representation of documents by word.

Índice

Índice de Figuras

Índice de Tabelas

Glossário de Termos

3M Terceiro Momento

ARFF *Attribute-Relation File Format*

ASCII *American Standard Code for Information Interchange*

CHI *Chi-Square*

CDD Classificação Decimal de *Dewey*

CDU Classificação Decimal Universal

FP Falsos Positivos

GR *Gain Ratio*

IG *Information Gain*

K-NN *K-Nearest Neighbor*

LCC *Library of Congress Classification*

MAT *Micro-Average Table*

MI *Mutual Information*

MicroR *Micro-Average Recall*

MicroP *Micro-Average Precision*

MF Matriz de Confusão

ROC *Receiver Operating Characteristic*

SVM *Support Vector Machines*

VP Verdadeiros Positivos

TF-IDF *Term Frequency Inverse Document Frequency*

R11 Subconjunto do *corpus Reuters* 21578 – 91 classes

R12 Subconjunto do *corpus Reuters* 21578 – 10 classes

R2 Subconjunto do *corpus Reuters* 21578 – Documentos multiclasse replicados

R3 Subconjunto do *corpus Reuters* 21578 – Documentos multiclasse considera-se a classe dominante

R4 Subconjunto do *corpus Reuters* 21578 – Excluíndo documentos multiclasse

Capítulo 1

Introdução

Classificar tematicamente um documento consiste em atribuí-lo a um ou mais temas, ou assuntos, pré-definidos. No processo de classificar um documento será necessário conseguir analisar o seu conteúdo e determinar, eficazmente, do que trata. Esta tarefa ainda hoje é realizada por profissionais dos diversos ramos do saber que analisam e interpretam cada documento.

Os bibliotecários, especialistas em interpretar os conteúdos temáticos em que a informação de um, ou de vários, documentos[1] se enquadram, utilizam como suporte diversos sistemas de catalogação. Designadamente o CDD (Classificação decimal de *Dewey*), LCC[2] (Classificação da Biblioteca do Congresso), e CDU (Classificação Decimal Universal). Devido à simplicidade da sua aplicação, e consulta, o CDU é o sistema mais utilizado a nível Europeu. O LCC encontra-se muito direccionado para os Estados Unidos (Valente s.d.).

Previamente à era dos computadores a extracção[3] do conteúdo dos documentos era realizada manualmente. Este trabalho consumia imenso tempo e também era muito susceptível a erros.

[1] Documentos neste contexto poderão ser: livros, cartas, revistas, artigos, entre outros.
[2] LCC também conhecido por *Library of Congress Classification*.
[3] Neste contexto "extracção" significa organizar o conteúdo do (s) documento (s) pelas dimensões que o compõem.

Käding, em 1987, e Thorndike, em 1921, são alguns dos autores que despenderam do seu tempo a realizar uma extracção manual em milhões de documentos[4] afim de verificarem a frequência de palavras, respectivamente alemão e inglês (Sardinha s.d.). Este trabalho teve por objectivo disponibilizar aos autores, principalmente de livros didácticos, uma base para produção de livros na língua materna. Tendo como maior preocupação auxiliar a produção de livros com a linguagem adequada a cada ano de escolarização. Contudo, trabalhos realizados desta forma acabaram por ser alvos de crítica, pois, contêm uma elevada margem de erro: numa extracção metódica é muito elevada a probabilidade de ocorrer um erro humano.

Quando se classifica computacionalmente um conjunto de documentos tem-se como objectivo melhorar e facilitar o acesso à informação. Para tal, será necessário representar computacionalmente o conteúdo dos mesmos. Só assim poderão ser interpretados por um classificador automático/computacional.

Um dos objectivos da automatização do processo de classificação é diminuir o trabalho do ser humano na atribuição de temas aos documentos a classificar. Com o auxílio do ser humano a máquina terá a capacidade de analisar o conteúdo dos documentos. Com o classificador escolhido e treinado pelo ser humano a máquina poderá atribuir o (s) tema (s) aos documentos a classificar. O ser humano terá de escolher os documentos de teste e, avaliar os resultados obtidos pelo classificador. Contudo, a atribuição dos temas é efectuado apenas pela máquina, ou seja, a classificação é automatizada[5].

Pretende-se conceber uma abordagem de Classificação Automática Supervisionada de documentos. O classificador antes de estar apto a classificar será treinado, e para tal serão utilizados documentos de treino. A classificação é supervisionada porque na fase de treino, se conhece a classe a que cada documento de treino pertence, e este conhecimento é utilizado no processo de treino do classificador. Por oposição, na classificação não Supervisionada, mais

[4] Käding - Extracção de cerca de 100 milhões de palavras da língua alemã (Käding s.d.).
Thorndike - Extracção de cerca de 625,000 palavras em literatura para crianças; 3,000 palavras da Bíblia, e livros de clássicos Ingleses; 300,000 palavras de livros da escola primária; 50,000 palavras de livros sobre: cozinha, costura, agricultura, entre outros; 90,000 palavras de jornais diários; e 500,000 palavras de cartas (Internet Archive s.d.).

[5] Automatizada no sentido de ter a capacidade de iniciar um processo, desenvolver e finalizar independentemente de ajuda exterior.

conhecida por agrupamento[6], não existe intervenção humana na indicação da classe. A classe é sugerida ou determinada pela abordagem, ou seja, emana apenas dos dados disponíveis e do uso do método de agrupamento. Apesar da vantagem de não necessitar que se indique a classe, tornando-o mais independente, tende a apresentar resultados mais modestos.

Na classificação dois conceitos muito utilizados são o *corpus* e o termo[7]. *Corpus* no contexto deste trabalho representa um conjunto de documentos que irão ser utilizados para treinar o classificador, *corpus* de treino[8]; e por *corpus* de teste, os documentos que vão ser classificados. Um termo representa cada uma das dimensões extraídas do *corpus* de treino. Poderá ser uma palavra, uma multi-palavra, uma cadeia de N caracteres, entre outros. Consoante a técnica de selecção de termos escolhida será analisada a frequência, simples ou pesada, a presença ou ausência, do termo nos documentos.

No processo de classificar computacionalmente existe uma fase que se denomina por pré-processamento de dados. Esta fase consiste em organizar computacionalmente o conteúdo dos documentos; decidir qual a representação a utilizar para o documento (palavras, multi-palavras ou cadeias de N-caracteres). Escolhendo no caso das palavras, e multi-palavras, como ocorrem no documento, ou normalizando-as para a sua forma canónica; finalmente, será necessário reduzir a dimensionalidade dos termos que se escolheu para representar os documentos.

A redução da dimensionalidade dos termos consiste em eliminar os termos que pouco, ou nada, contribuem para a classificação. Trata-se de termos que não contêm um peso significativo para o tema que se pretende classificar. Nos capítulos 1.1. e 2.2 será explicado com maior pormenor o significado deste peso.

A redução da dimensão dos termos não é indispensável para a classificação. Existem classificadores, como o RIPPER (Cohen e Singer s.d.), que utilizam a representação íntegral[9] dos

[6] Do termo inglês, *clustering*.
[7] Do Inglês, *feature*.
[8] C*orpus* de treino trata-se de um conjunto de documentos utilizados para treinar o classificador.
[9] Representação íntegral - todos os termos são utilizados por o classificador, ou seja, não existe uma redução de dimensionalidade prévia à classificação.

documentos. No entanto, se não for reduzida a dimensionalidade dos termos, a performance computacional de outros classificadores, como o KNN (Wikipedia s.d.), torna-se inexequível.

Antes de introduzir os termos no classificador poder-se-ão utilizar técnicas de selecção de termos, ou de redução de dimensionalidade da representação dos documentos. Estas técnicas vão analisar os termos e pontuá-los conforme a sua importância no (s) tema (s) a classificar. Com os resultados obtidos poder-se-ão excluir os termos com pontuação menor pois estes só vão causar ruído[10]. Estas técnicas vão ajudar a melhorar o desempenho do classificador.

Durante este trabalho utilizei diferentes classificadores e diferentes técnicas de selecção de termos. Os resultados obtidos são objecto de avaliação:

1. Os motivos porque uma técnica demonstra resultados similares, ou diferentes, em relação a outra (s) técnica (s).
2. O tempo de processamento.
3. A selecção dos termos que melhor identificam os temas a classificar.

Perante um fenómeno novo um ser humano visualiza, interpreta e raciocina sem ser forçosamente necessário ensinar algo novo sempre que este surge. As regras das técnicas de aprendizagem para uma máquina, são muito distintas das do ser humano, e assim, ir-se-á observar como a máquina irá "reagir" à aplicação de diferentes técnicas de selecção de termos, e utilizando mais do que um classificador. Neste trabalho analisarei, e concluirei, sobre a qualidade dos resultados obtidos.

[10] Do Inglês, *noise*.

1.1. Motivação

Nos dias de hoje a classificação automática de documentos tornou-se indispensável. Técnicas como CDD, CDU e LCC, exigem um consumo muito elevado do tempo de trabalho de especialistas altamente qualificados. E seria impossível acompanhar o ritmo a que surgem novos documentos se se utilizassem apenas técnicas de classificação tão morosas. Se o trabalho da análise do conteúdo da informação for realizado por uma máquina, os resultados são obtidos muito mais rapidamente, ainda que tenham de ser certificados à *posteriori*.

A questão do tempo gasto a classificar cada documento é determinante. Classificadores automáticos que não suportam uma quantidade de documentos muito elevada tornam-se prescindíveis em algumas áreas, como por exemplo, classificação automática de documentos na Internet. O tempo de computação a classificar poderá ser um factor de exclusão na utilização de alguns classificadores que sejam muito morosos.

A tendência de a informação ser convertida, ou apenas existir, em formato digital tem vindo a aumentar. Um dos exemplos desta tendência é o aparecimento de várias bibliotecas digitais: A Biblioteca Europeia (The European Library s.d.); Biblioteca Nacional Digital Brasil (Biblioteca Digital do Brasil s.d.); entre muitas outras. Outrora para se poder consultar a informação contida numa biblioteca seria necessário deslocarmo-nos às suas instalações. Este avanço dos tempos ajudou no processo de classificar automaticamente: a informação encontra-se em formato digital. Assim, esta tendência incentiva a necessidade de estudar como os classificadores operam, como melhorar a sua performance; como superar as suas limitações; entre muitas outras motivações. Trata-se de uma área que ainda não estagnou, e, por isso, existe a necessidade de a estudar e, se for possível, melhorar as técnicas existentes.

Neste trabalho pretende-se analisar automaticamente documentos, e, finalmente classificá-los. Para esse efeito é necessário definir como representar cada documento de forma a poder ser trabalhado computacionalmente. A representação mais usual é considerar o texto como um vector de palavras, em que cada palavra define uma dimensão no vector. Mas também se pode considerar outro tipo de dimensões: multi-palavras; caracteres; ou cadeias de caracteres (com

tamanho definido no âmbito do trabalho). A representação por caracteres tem interesse em estudos que tenham por objectivo realizar uma identificação da língua em que o documento está escrito, pois existem línguas, em que alguns caracteres se evidenciam com maior frequência (Silva, et al. s.d.). Pretende-se com este trabalho categorizar tematicamente, e para tal utilizou-se mais do que um tipo de dimensão. Só assim será possível comparar os diferentes resultados obtidos: a escolha do tipo de dimensão/termo utilizado poderá ter um peso relevante no desempenho de alguns classificadores.

Com os avanços contínuos da capacidade de armazenamento e processamento dos computadores, não é impossível introduzir no classificador a representação íntegra de cada documento como acontece, por exemplo, com o classificador RIPPER (Cohen e Singer s.d.). Contudo, trata-se de uma quantidade muito elevada de termos, onde muitos pouco, ou nada, contribuem para a classificação. Daí a necessidade de se proceder a uma selecção prévia dos termos mais relevantes.

Os termos utilizados pelos algoritmos de classificação podem incluir parcialmente, ou por completo (considerando que nem sempre são utilizadas palavras), nomes de pessoas, lugares, marcadores temporais, acontecimentos ou raízes de palavras que se comprovem relevantes para efeitos de discriminação. Compete à técnica de selecção de termos determinar esse poder discriminante. A decisão de seleccionar, ou eliminar, os termos mais relevantes, ou irrelevantes é uma etapa determinante nas técnicas de *Machine Learning*[11] (Blum e Langley s.d.).

Os termos que não são discriminantes para o (s) tema (s) a classificar aparecem usualmente com uma elevada frequência no (s) documento (s). Estes termos deverão ser excluídos pois apenas deterioram o desempenho do classificador. As técnicas de selecção dos termos mais relevantes têm a função de determinar quais os termos que serão excluídos do processo de classificação.

A selecção dos termos mais relevantes é, em grande parte, o segredo da classificação. Estes termos, por vezes, não são os que se repetem mais frequentemente. Deverão, contudo, ser os mais representativos. É importante reduzir a dimensionalidade do número de termos contidos em cada documento, mantendo, apenas, os que melhor representam o seu conteúdo temático. Um termo

[11] *Machine Learning* está englobado numa subárea de inteligência artificial, em que se desenvolvem algoritmos e técnicas que permitem que a máquina "aprenda".

que se manifeste em todos os documentos está a indicar que não demonstra nenhum género de conexão com os temas (considerando um *corpus* de diferentes temas). A sua frequência relativa, em relação à sua presença em outros documentos, é muito elevada o que leva a diminuir a sua contribuição para a classificação.

Exemplo:

1. "Economia é a ciência social que estuda a produção, distribuição, e consumo de bens e serviços. (Wikipedia s.d.)"
2. "Há evidências de que a música é conhecida e praticada desde a pré-história. (Wikipedia s.d.)"

O artigo "a", o pronome "que", a conjunção "e", e a forma "é" do verbo "ser" apresentam-se em ambas as frases. No entanto nenhum é característico do tema Economia ou Música. Será importante que os resultados obtidos pelo classificador não sejam influenciados pela presença de termos como estes.

Pode-se, assim, indicar que o trabalho realizado se encontra dividido em cinco etapas:

1. Definir a representação segundo a qual se pretende estruturar os documentos para que possam ser interpretados computacionalmente.
2. Definir o conjunto de termos com maior relevância
3. Escolher, e treinar o classificador com o *corpus* de treino.
4. Classificar novos *corpus* (de teste).
5. Avaliar os resultados obtidos com o classificador.

Com este trabalho pretende-se trabalhar três classificadores, com três métodos de selecção de termos, e sete representações de documentos. Os resultados das diferentes combinações (representação de documento/método de selecção de termos/classificador) serão objecto de análise, e avaliação, quer quanto ao seu desempenho temporal, quer relativamente à precisão obtida.

É na quinta etapa que se identificarão as técnicas de selecção de termos com as quais se obtêm resultados mais fiáveis; identificar-se-á também, se a utilização de diferentes tipos de dimensões

influem nos resultados obtidos; entre outros factores. É nesta etapa que se poderá observar os resultados obtidos através da utilização de diferentes tipos de dimensões, diferentes técnicas de selecção de termos, e diferentes classificadores.

De forma a ser possível realizar a análise comparativa dos resultados obtidos recorreu-se a técnicas de classificação já experimentadas, e comprovada a sua aplicabilidade na área de categorização temática. No capítulo 1.2 será brevemente indicado quais os classificadores a utilizar.

Das várias técnicas existentes para selecção de termos existem algumas que apresentam um bom desempenho. Mas apesar dos seus bons resultados ainda há aspectos que necessitam de ser melhorados (Sebastiani, Machine Learning in Automated Text Categorization s.d.).

Pretende-se implementar e comparar os resultados obtidos com uma nova técnica de selecção de termos: Terceiro Momento centrado em relação à média. Até ao momento ainda não foi realizado nenhum estudo comparativo com as técnicas mais conhecidas, e muito utilizadas. Técnicas que no próximo capítulo irão ser descritas: o seu funcionamento, e o que mais valorizam na atribuição de pesos aos termos.

Com esta comparação pretende-se determinar quais as vantagens, e desvantagens, de aplicar o terceiro momento em alternativa a outra técnica. Considera-se fundamental a realização deste trabalho pois existem muitas técnicas de selecção de termos que apresentam bons resultados, mas será deveras importante determinar se o terceiro momento será uma técnica que vai prevalecer sobre outras técnicas já existentes.

No capítulo 3 são apresentados os resultados obtidos com os três classificadores, três técnicas de selecção de termos e as sete representações de documentos utilizadas. Observa-se que a técnica do terceiro momento tem a capacidade de seleccionar termos com um desempenho equivalente ao das restantes técnicas. Sendo, por vezes, a técnica que consegue retornar o melhor desempenho (ver resultados do classificador SVM na tabela 3.7.1.).

1.2. Solução apresentada

No trabalho realizado foram experimentados sete tipos de termos/dimensões de representação de documentos, três técnicas de selecção de termos, e finalmente, três classificadores. Pretende-se com o mesmo classificador observar os resultados obtidos com as diferentes técnicas de selecção, e diferentes tipos de dimensões. As dimensões escolhidas são: a palavra, cadeias dos primeiros quatro, cinco, seis caracteres individualmente, e globalmente; pentagramas[12]; palavras; e multi-palavras[13]. Os pentagramas diferem da representação pelo primeiros 5 caracteres, na medida em que a representação dos primeiros 5 caracteres apenas tem em consideração, como termo, os primeiros cinco caracteres de uma palavra. Por pentagramas uma palavra, se for de tamanho superior ou igual a seis, irá corresponder a mais do que um termo. Por exemplo a palavra "medida", tem seis caracteres, na representação com os primeiros 5 caracteres corresponderá a um único termo: "medid", mas por pentagramas obter-se-ão dois termos: "medid" e "edida".

Como já foi referido previamente, a medida do Terceiro Momento foi recentemente proposta, e por isso, não existia trabalho de comparação realizado sobre o seu desempenho em relação a outras medidas já existentes, e experimentadas. Logo, realizou-se uma experimentação com diferentes conjuntos de dados, utilizando diferentes medidas, ou seja, diferentes metodologias de atribuição de pesos aos termos. Entre estas medidas estará o Terceiro Momento.

A medida de selecção de termos *Chi-square* (Debole e Sebastiani s.d.) (Yiming e Jan s.d.) e *Information Gain* (Debole e Sebastiani s.d.) (Yiming e Jan s.d.) (Mladenic s.d.) (Krkoska, Pekar e Staab s.d.) (Hall e Holmes s.d.) foram as medidas escolhidas para comparação com o Terceiro Momento. Estas medidas já foram utilizadas por vários autores em diferentes contextos e na maioria dos casos apresentaram bons resultados. No próximo capítulo explicar-se-á de forma mais extensa como funcionam estas medidas. Para servir de termo de comparação reportam-se os

[12] Entende-se por pentagramas como combinações de cinco caracteres por palavra. No caso da palavra classificar os pentagramas seriam: class; lassi; assif; ssifi; sific; ifica; e ficar.
[13] Multi-palavra corresponde a combinações de palavras. Por exemplo: câmara municipal; câmara escura; câmara alta; câmara legislativa; entre outras.

resultados obtidos com o classificador RIPPER em que não é utilizada qualquer medida de selecção de termos.

Escolhidas as medidas de selecção de termos foi necessário escolher qual o classificador a utilizar. Cada uma das medidas de selecção de termos poderá originar diferentes resultados conforme o classificador escolhido. Se tivesse sido empregue apenas um classificador, a respectiva conclusão sobre a performance da medida do terceiro momento em relação às outras medidas poderia ser deturpada. Isto acontece porque um classificador poderá retornar resultados mais favoráveis aquando da utilização de uma técnica, e o contrário também poderá acontecer, e assim, adulterar as conclusões sobre o seu desempenho. Com outros classificadores os resultados poderão ser menos optimistas, ou pessimistas. Para evitar tal situação foram escolhidos três classificadores, previamente estudados, que apresentam bons resultados e uma boa performance, demonstrando resultados fiáveis em diferentes aplicações. Estes classificadores foram *SVM* (*Suport Vector Machines*), K-NN (*K-Nearest Neighbour*) e *RIPPER*.

Os classificadores *SVM* (Sebastiani, Machine Learning in Automated Text Categorization s.d.) (Thorsten n.d.), K-NN (Sebastiani, Machine Learning in Automated Text Categorization s.d.) (Thorsten n.d.) e *RIPPER* (Sebastiani, Machine Learning in Automated Text Categorization s.d.) (Cohen e Singer s.d.) estão disponíveis na ferramenta *WEKA* (Hall e Holmes s.d.). E por isso neste trabalho *WEKA* foi uma ferramenta auxiliar que ajudou no desenvolvimento deste estudo.

A técnica de selecção de termos, Terceiro Momento, como foi referido, ainda não foi alvo de nenhum estudo sobre o seu desempenho. Esta técnica visa detectar as características mais relevantes com o auxílio de uma métrica baseada no conceito estatístico do terceiro momento centrado em relação à média. Esta opção emerge da necessidade de aproveitar a capacidade deste conceito em detectar *outliers*[14] (Wikipedia s.d.), ou seja, de detectar e promover os elementos que ocorrem fora da norma[15], e em especial o elemento que é invulgar. E assim, com o Terceiro Momento é possível tirar partido desta propriedade das funções cúbicas.

[14] Denomina-se por *outlier* o termo que numericamente se afasta dos elementos do conjunto.
[15] Neste contexto entende-se como norma o valor médio da frequência que cada documento apresenta para uma determinada dimensão (sequência de caracteres ou palavras).

Na expressão 1.1 pode-se observar como se determina o poder discriminante de um termo t utilizando o terceiro momento:

$$V(t) = \frac{\frac{1}{C.\,p_m(t,.)}\sum_{c=1}^{c=C}[p_m(t, classe_c) - p_m(t,.)]^3}{Disp_m(t)}$$

C: Número total de classes .
$p_m(t, c)$: Probabilidade média de um termo ocorrer na classe c.
$p_m(t, .)$: Probabilidade média de um termo ocorrer em todas as classes do *corpus*.
$disp_m(t)$: Dispersão[16] média do termo nas classes

(1.1)

No capítulo 2.2.6. estará uma descrição mais extensa da expressão 1.1. e dos seus componentes.

É importante saber quais os termos que são invulgares, assim como o sentido de afastamento destes termos. Se este fôr positivo, significa que possivelmente se trata de um termo característico da classe - informação preciosa na atribuição de um maior peso aos termos mais relevantes[17]; se o afastamento for negativo não será uma informação muito relevante. Apenas poderá indicar que o termo é inexistente nalguma classe ou surge invulgarmente com menor peso nela.

Para avaliar o poder discriminante de um termo, esta métrica terá em consideração dois aspectos:

1. A variação da probabilidade média de ocorrência do termo nas diferentes classes, que deverá ser positiva por ser resultante da existência dum afastamento positivo, quando o termo é discriminante;
2. A dispersão média, por classe, da probabilidade de ocorrência do termo nos documentos dentro da mesma classe.

[16] Disperso no sentido de ter uma variação elevada.
[17] Relevantes no sentido de serem mais específicas/fiéis à classe.

Exemplo:

Suponhamos a existência de três classes (Agricultura, Energia e Ciência) e um termo y.

A frequência do termo nos documentos poderá apresentar a seguinte distribuição:

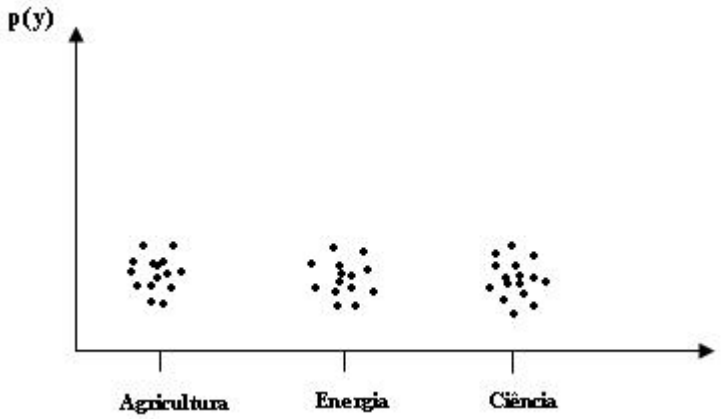

Figura 1.2. 1: Exemplo Ilustrativo 1

Observando a Fig.1.2.1, o termo y, intuitivamente, não é determinante para a discriminação das classes Agricultura/Energia/Ciência, pois apresenta-se com distribuições semelhantes em todas as classes. Parece tratar-se de um termo comum a todos os documentos, mas não determinante para a classificação da classe.

Suponhamos agora que um termo z apresente as seguintes distribuições:

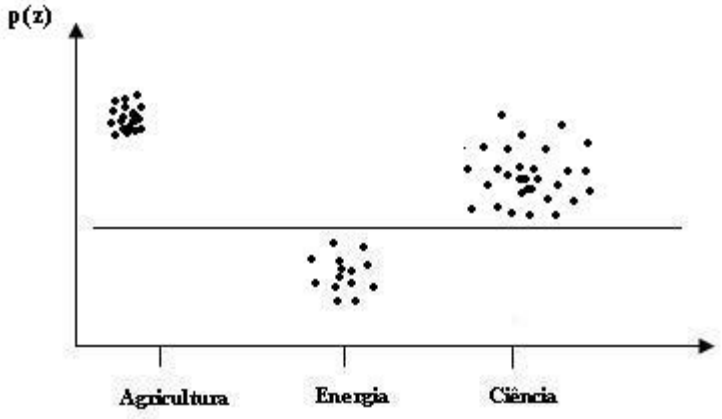

Figura 1.2. 2: Exemplo Ilustrativo 2

Observa-se que o termo z é, possivelmente, um termo característico da classe Agricultura já que, por um lado, a probabilidade média de z nesta classe é elevada, e por outro a sua dispersão nela é muito pequena. O termo z parece ser «fiel» à classe Agricultura e portanto característico dela. De notar que, o termo z embora surja na classe Ciência com probabilidades acima da média (denotada pela linha horizontal), a sua dispersão dentro desta classe, por ser relativamente elevada, impõe-lhe um peso baixo na sua caracterização.

Por fim facilmente se compreende que na Fig. 1.2.2. z não é um termo característico da classe Energia.

1.3. Principais contribuições

Nesta tese inovei ao introduzir uma nova técnica de selecção de termos (3º momento em relação à média) e ao ter comparado essa técnica com outras técnicas já experimentadas anteriormente (*Chi Square* e *Information Gain*).

Neste momento tem-se conhecimento do desempenho da medida do Terceiro Momento em relação a medidas já muito conhecidas, e recomendadas como técnicas que apresentam um bom desempenho. E pode-se confirmar que se trata de uma medida que tem a capacidade de seleccionar termos que discriminam o *corpus* com diferentes representações computacionais.

Inovei ainda ao experimentar diversas formas de representação dos textos para além da mais usual em trabalhos deste tipo (com base em palavras, com ou sem filtragem de *stopwords* e recorrendo a lematização). Verifiquei que os resultados da classificação obtidos utilizando os primeiros 4 caracteres das palavras se aproximavam muito dos obtidos com palavras o que torna esta técnica que utilizei ainda mais independente das línguas na medida em que não preciso de lematizadores nem de listas de *stopwords*.

Capítulo 2

Trabalho relacionado

Este capítulo analisará trabalhos relacionados com o tema central desta tese.

2.1. Representação computacional dos documentos

No âmbito deste trabalho, a representação dos documentos de treino, e dos documentos a classificar, poderá ser realizada utilizando:

a. Palavras - Cada dimensão dos documentos corresponderá a uma palavra.

b. Sequência de N palavras (N-gramas de palavras) - Cada dimensão será uma sequência de N palavras, com N igual a 2,3,etc. Um termo poderá ser composto por duas, três, quatro palavras, e assim consecutivamente. Em trabalhos realizados com sequências de palavras foi observado que não se melhoravam os resultados com o uso de quatro palavras (Papka e Allan s.d.) (Jacquemin, Klavans e Tzoukermann s.d.).

Em particular podem ser utilizados termos com mais do que uma palavra que tenham associados a si significados precisos, por exemplo:

- Câmara Escura
- Câmara Municipal
- Câmara Alta
- Câmara de Filmar

É importante que a extracção deste tipo de termos compostos seja automática (Silva, et al. s.d.).

c. Cadeia de N-caracteres[18] - Cada dimensão será uma cadeia de N caracteres com um valor inteiro superior a um. Neste trabalho pretende-se categorizar tematicamente documentos e caracteres isolados não acarretam informação temática relevante. Línguas como as Orientais são uma excepção a esta característica, e por esta razão, serão excluídas do trabalho que se realizou.

d. Multi-Palavra – Uma multi-palavra representa uma sequência de N-gramas[19] de palavras, todavia, uma multi-palavra é um conjunto de palavras que se deseja que tenham necessariamente um significado. Estas palavras ou são extraídas tendo em linha de conta informação morfossintática de cada um dos seus constituintes, não sendo por isso a sua extracção independente da língua (19), ou são extraídas tendo em linha de conta o grau de coesão estatística entre os seus constituintes (20) e (Aires, Lopes e Silva s.d.), sendo neste caso a sua extracção independente da língua. Por exemplo, no texto:
    - "A Câmara Municipal de Cascais organiza a segunda edição do Concurso de Fotografia Digital."
        - São sequências de 2-gramas de palavras:
          A Câmara; Câmara Municipal; de Cascais; Cascais organiza; organiza a; a segunda; segunda edição; edição do; do Concurso; Concurso de; de Fotografia; e Fotografia Digital.

[18] Do termo N-grams da língua inglesa.
[19] N-gramas corresponde a uma sequência de N letras ou N palavras. Para a representação computacional por multi-palavra será N palavras.

- São Multi-palavras:
  Câmara Municipal; Câmara Municipal de Cascais; Concurso de Fotografia; Fotografia Digital; e Concurso de Fotografia Digital.

Na realização deste estudo utilizou-se a representação com base em palavras, cadeias de caracteres, e finalmente, com multi-palavras extraídas automaticamente do *corpus* utilizando (Aires, Lopes e Silva s.d.).

2.2. Redução da Dimensionalidade do *Corpus*

Existem muitas formas de remover os termos/dimensões que menos importam ao processo de treino de classificadores, e de classificações de novos documentos. Com esta redução de dimensionalidade pretende-se dar maior importância a termos/dimensões que se encontravam camuflados devido à elevada frequência de termos/dimensões não-informativos. Por exemplo, os artigos, as preposições, as conjunções ou pronomes, existem em qualquer documento de língua portuguesa, e em outras línguas, e contêm frequências relativas muito elevadas. Contudo, palavras deste tipo não são informativas do assunto predominante no documento, pelo que é usual não as considerar, em particular no processo de classificação temática de documentos. Estas são palavras que usualmente são conhecidas como *stop-words*[20].

Muitos autores como (Pons-Porrata, Berlanga-Llavori e Ruiz-Shulcloper s.d.), (Kumaran e Allan s.d.), (Galho e Moraes s.d.), (Wong e Fu s.d.), (Nallapati s.d.), e (Mathieu, Besancon e Fluhr s.d.) optaram por uma utilização de listas de palavras não interessantes (*stop-words*) para as eliminar do processo de treino e de classificação.

Apesar de ser uma técnica que consegue reduzir a dimensionalidade do vector de dimensões, removendo informação não-relevante para a classificação, trata-se de uma técnica muito estática porque não identifica automaticamente as *stop-words* e não evolui, já que não aprende. Além disso, é dependente da língua em causa, visto as *stop-words* (tipicamente artigos, preposições, conjunções, etc.**)** variarem de língua para língua.

Existem outras técnicas que diminuem o número de termos, ou dimensões. Duas das mais utilizadas são a radicalização[21] (Wikipedia s.d.) e a lematização[22] (Wikipedia s.d.), que têm por objectivo representar palavras da mesma família por uma única, acumulando assim a frequência das diferentes formas de palavras relacionadas. Este processo substitui cada palavra pelo seu

[20] É o nome atribuído às palavras que, por «não terem conteúdo semântico», são filtradas dos textos, para vários efeitos. Tipicamente são palavras com elevada frequência nos documentos.
[21] Do inglês, *stemming*.
[22] Do inglês, *lemmatization*.

radical (*stem*) ou lema (*lemma*), ou seja, por um seu representante. O radical de uma palavra é a parte da palavra que é comum a todas as que pertencem à mesma família (ex. espera, esperar, esperaram, esperança, desespero, desesperança); e lematização consiste na transformação duma palavra na sua forma canónica (lema). Isto é, os tempos verbais são passados para a forma infinitiva, as formas adjectivais passam a ser representadas pela forma singular masculina, os nomes no plural pelas respectivas formas no singular, etc.

Existem variantes deste processo, de radicalização e lematização, como: no artigo (Pons-Porrata, Berlanga-Llavori e Ruiz-Shulcloper s.d.) a etiquetagem morfossintática[23] (Wikipedia s.d.), no artigo (Nallapati s.d.) (Bikel, Schwartz e Weischedel s.d.) (Xu, Broglio e Croft s.d.) o *Identifinder*[24], e no artigo (Kumaran e Allan s.d.), (Broglio, Callan e Croft s.d.) o algoritmo *K-stem* e *InQuery*.

A técnica de etiquetagem morfossintática surge para auxiliar o processo de lematização e radicalização. Associa cada palavra à sua classe morfossintática, ajudando assim no processo de desambiguação de qual a forma canónica a adoptar.

Por exemplo, a palavra guarda nos exemplos seguintes:

1. "Hoje estive a passear na **Guarda**."
2. "**Guardas** as cartas na cómoda."
3. "O **guarda** apanhou o ladrão."
4. "Ele tem um cão de **guarda** de raça pura."

é um substantivo, nome próprio (localidade), forma do verbo guardar. É possível observar que se trata de uma palavra que contém significados diferentes de acordo com os contextos em que está inserida.

No processamento por computador, para a língua inglesa, a etiquetagem morfossintática distingue entre 50 a 150 etiquetas (Wikipedia s.d.). Com o aparecimento de novas técnicas de desambiguação de termos, e implementação de algoritmos dinâmicos, na década de 80, é que esta

[23] Do Inglês, *part-of-speech tagging*.
[24] Ferramenta que identifica as *named-entities*, ou entidades nome, (termos que referem pessoas, organizações, locais, expressões temporais ou expressões numéricas).

técnica foi considerável viável para ser utilizada no âmbito de mineração de dados[25]. Para a língua portuguesa foi utilizado por (Marques e Lopes s.d.) um etiquetador treinável baseado em redes neuronais, em que foram utilizadas 40 etiquetas, demonstrando uma boa, e rápida, performance.

O *Identifinder* trata-se de um modelo que consiste em identificar, e classificar dentro de um grupo (Pessoa, Organização, não-Nome), termos que possam representar: localidades; nomes pessoais; datas; montantes monetários; percentagens; entre outros. Nesta técnica é tida em consideração a relação entre um termo e o termo que o precede no documento (Bikel, Schwartz e Weischedel s.d.). Ao se identificar os termos como: localidades, nomes, locais, entre outros grupos, está-se a ajudar no processo de redução de dimensionalidade. Usualmente estes termos não se repetem muito, no entanto poderão ser característicos da classe. Por exemplo, o nome Fabrizio Sebastiani (Sebastiani, A Tutorial on Automated Text Categorization s.d.) seria discriminante para o tema "Classificação de Documentos".

InQuery é um sistema implementado em *UNIX* que disponibiliza várias técnicas de indexação de termos:

1. Indexação por palavras;
2. Indexação por *part-of-speech tagging*;
3. Indexação por dependência do domínio dos termos (datas, locais, nomes de companhias, entre outros).

InQuery reduz a dimensionalidade das dimensões por indexar; remover *stop-words*, identificar etiquetagem morfossintática; identificar o domínio de termos; entre muitos outros. No artigo (Broglio, Callan e Croft s.d.) encontra-se descrito com maior pormenor como esta ferramenta funciona.

Também se verifica que estas técnicas (etiquetagem morfossintática, *Identifier*, *InQuery*) são limitadas a nível de uso que se deseje que seja independente da língua, pois toda a informação é dependente da língua em questão, e não estão disponíveis para qualquer língua.

[25] Do Inglês, Data Mining.

Um dos objectivos mais importantes nesta fase de redução da dimensionalidade dos documentos é o de conseguir obter um método capaz de avaliar quais as candidatas a serem descartadas, e que não seja muito dispendioso a nível computacional.

As técnicas de selecção de termos podem ser realizadas de duas formas (How e Narayanan s.d.):

1. Localizada:
   Um conjunto de candidatas (termos) é definido com base na informação relevante e não-relevante em documentos para uma classe. Logo, cada classe é representada por um conjunto único de termos.
2. Global:
   Nestas técnicas é considerada a hipótese de os termos serem partilhados por mais do que uma classe.

Existem também outras duas variantes nas medidas de selecção de termos, que contemplam quer pesos negativos quer positivos, ou seja, o quão próximo uma candidata está para uma classe como também o seu afastamento. Outra hipótese será apenas considerar os termos que discriminam positivamente uma classe sem ter em consideração termos de outras classes (afastamento negativo).

Usualmente estabelece-se um limiar de corte para se considerar apenas os valores que serão significativos para a classificação. Um termo que seja pontuado negativamente, ou positivamente, apenas será avaliado pelo seu valor absoluto[26]. Assim é possível eliminar os termos que pouco contribuem para a discriminação dos temas a classificar.

Finalmente entre as variantes de técnicas de selecção de termos acima descritas é sempre necessário decidir se a opção é reduzir a dimensionalidade do espaço de termos por extracção ou por selecção de termos. Se for por selecção de termos será escolhido um subconjunto de termos do conjunto inicial. Se for por extracção de termos, os termos extraídos não serão um subconjunto do conjunto inicial, mas sim combinações ou transformações dos termos iniciais

[26] Valor absoluto no sentido de não se considerar se é positivo ou negativo. Pretende-se saber o quão próximo ou afastado o termo se encontra.

(Sebastiani, A Tutorial on Automated Text Categorization s.d.). Também é possível utilizar ambas as técnicas (Silva, et al. s.d.), selecção e extracção.

Alguns dos métodos mais comuns para determinar a relevância de uma dimensão num documento serão descritos nas próximas secções.

2.2.1. *Term Frequency*

A técnica TF (*term frequency*) indica o número de vezes que um termo ocorre num documento. Assim, os termos que ocorrem com maior frequência são considerados mais relevantes do que os menos frequentes.

No entanto, um termo pode ser muito frequente, não só num conjunto de documentos, como no domínio todo. Acontece que, se um termo ocorre no conjunto todo, não deve ser tomado em conta como relevante uma vez que não é característico. Contudo a técnica TF, sem se combinar com outras técnicas, apenas contabiliza a ocorrência do termo, dando-lhe uma relevância excessiva, tendo em conta a sua importância real no documento.

Esta medida assume que os termos raros não são informativos para a previsão da categoria, ou que, devido à sua baixa frequência, não influenciam a performance global do classificador que irá funcionar com base nos termos escolhidos. Assim, a remoção destes termos reduz a dimensionalidade do espaço de termos de uma forma simples e de baixo gasto de tempo computacional, muito próximo do linear. No entanto, muitas vezes estes termos rejeitados poderiam melhorar a performance da classificação (Yiming e Jan s.d.). Os termos mais representativos de um tema têm uma maior probabilidade de ocorrer com uma frequência menor do que termos comuns a temas diferentes. Assim, ao excluir estes termos perder-se-á informação vital para a performance do classificador. Claro que existe sempre a possibilidade de seleccionar apenas os termos que não façam parte do conjunto de termos que cumulativamente ocupem, por exemplo, 40% do *corpus* de teste. Esta medida evitava assim os termos muito frequentes, deixando apenas os menos frequentes.

2.2.2. *Relative Frequency*

Na técnica RF (*Relative frequency*) o valor de TF é normalizador visto dividir a frequência do termo pelo número total de termos no documento. Assim, é possível obter a mesma ordem de grandeza, ou seja, relativizar de forma comparável, o mesmo termo em documentos diferentes porque todos os documentos vão ficar com um valor de RF (entre um e zero). No entanto continuamos com o problema dos termos, que ocorrem em muitos documentos, terem relevância equivalente aos termos que ocorrem em poucos documentos.

2.2.3. *Inverse Document Frequency*

IDF (*Inverse document frequency*) é uma medida que é usada para penalizar os termos que ocorrem em muitos documentos de classes diferentes. Favorece termos que aparecem pouco em documentos distribuídos por poucas classes. Assim, apesar de um termo poder ser muito frequente num documento, este cálculo tem em conta a dispersão do termo dentro do conjunto de documentos. Os termos «exclusivos» de um documento são valorizados. Esta valorização é alcançada calculando o TF pesado pela dispersão do termo pelos documentos.

Mais precisamente, para cada termo temos:

$$IDF = log \frac{N}{C} \tag{2.1}$$

N representa o número total dos documentos, e C representa o número de documentos que contem o termo.

De notar que esta métrica é insensível à distribuição das ocorrências pelos diferentes documentos. Além do mais valoriza excessivamente as ocorridas por lapsos ortográficos.

2.2.4. *Term Frequency Inverse Document Frequency*

TFIDF (*Term frequency inverse document frequency*) é uma medida que determina o número de vezes que o termo ocorre num documento (TF) e penaliza os termos que ocorrem em muitos documentos (IDF).
Sendo possível comparar entre documentos diferentes os pesos obtidos para cada termo, pois estão na mesma ordem de grandeza (RF).

$$TFIDF = TF * IDF \tag{2.2}$$

2.2.5. *Chi-square*

Chi-square, χ^2, é uma medida que se baseia num método probabilístico que interpreta um evento num conjunto de documentos, calculando desta forma a probabilidade de um termo característico de uma classe, pertencer ou não pertencer a um documento aleatoriamente escolhido (Debole e Sebastiani s.d.) (Hashimoto e Yukawa s.d.). Assim, *Chi-square* mede a falta de independência entre um termo t e uma classe c:

$$\chi^2(t,c) = \frac{N.(AD - CB)^2}{(A+C).(B+D).(A+B).(C+D)} \tag{2.3}$$

A representa o número de vezes que o termo t e a classe c co-ocorrem;
B representa o número de vezes que o termo t ocorre sem ser na classe c;
C representa o número de vezes que a classe c ocorre sem o termo t;
D representa o número de vezes que nem a classe c, nem o termo t ocorrem; e
N representa o número total de documentos.

No cálculo dos pesos dos termos utilizando a medida Chi-square os termos que pesam negativamente para uma classe não são ignorados. A medida calcula a frequência da presença e da ausência de um termo numa classe.

Esta medida é muito usada em experiências científicas com o fim de se observar o quão diferente serão os resultados obtidos dos resultados esperados (de acordo com uma hipótese inicial) (Debole e Sebastiani s.d.).

2.2.6. *Odds ratio*

Odds-ratio é uma medida probabilística que categoriza os termos de acordo com a sua relevância numa associação com conjuntos de termos negativos, e um conjunto de termos positivos em relação a uma classe: um termo pode afastar-se negativamente, ou aproximar-se positivamente de uma classe.

Se fôr um termo positivo, e quanto maior fôr o seu valor, indica que será mais discriminante para a classe. Se for negativo indica que se trata de um termo que se afasta da classe em questão.

Trata-se de um método bem sucedido na sua capacidade de seleccionar características para classificar documentos (Mladenic s.d.).
Dado um termo t, num conjunto de exemplos positivos e negativos de uma classe c (Krkoska, Pekar e Staab s.d.):

$$OR(t,c) = \frac{p(t|c).\left(1 - p(t|\bar{c})\right)}{\left(1 - p(t|c)\right).p(t|\bar{c})} \tag{2.4}$$

$p(t|c)$: A probabilidade de um termo t ocorrer dado que o documento pertence à classe c.
$p(t|\bar{c})$: A probabilidade de um termo t ocorrer dado que o documento não pertence à classe c.

Com esta medida tem-se em consideração a diferença da distribuição dos termos nos documentos que são relevantes, para os temas a classificar, em relação aos documentos que não são relevantes.

2.2.7. *Information Gain*

IG (*Information Gain*) é uma medida que prevê a categoria pela indicação da presença ou ausência de um termo no documento (Yiming e Jan s.d.).

Considerando que t representa um termo e c uma categoria, as seguintes expressões indicam como se observa o valor da entropia[27] da categoria antes, e depois de observar o termo (Hall e Holmes s.d.):

$$H(c) = -\sum_{c \in C} p(c) \log_2 p(c) \tag{2.5}$$

$$H(c|t) = -\sum_{t \in T} p(t) \sum_{c \in C} p(c|t) \log_2 p(c|t) \tag{2.6}$$

$p(t)$: A probabilidade de um termo t ocorrer.
$p(c)$: A probabilidade de uma classe c ocorrer.
$p(c|t)$: A probabilidade de uma classe c ocorrer dado que ocorre o termo t.

Se o valor da entropia diminuir, indica que o termo poderá ser indicativo para a classe. Se esta não se encontrar muito dispersa na classe é porque provavelmente é característica dessa classe.

Na medida IG é aproveitada a informação da entropia do termo. Para determinar o peso de cada termo é-lhe atribuído uma pontuação tendo em consideração o valor da entropia da classe e da entropia do termo *i* na classe:

$$\begin{aligned} IG_i &= H(C) - H(C|T_i) \\ &= H(T_i) - H(T_i|C) \\ &= H(T_i) + H(C) - H(T_i, C) \end{aligned} \tag{2.7}$$

Assim pode-se afirmar que IG mede a quantidade de informação que um termo escolhido ao acaso contém em relação a outro termo. Se forem dois termos aleatoriamente independentes

[27] Neste contexto entropia representa a dispersão de um elemento numa dada classe.

retorna 0, senão cresce monotonicamente com a dependência entre os termos (Debole e Sebastiani s.d.).

2.2.8. *Gain Ratio*

Gain Ration é uma versão normalizada da medida IG. Como se pode observar na definição (2.7) IG cresce com a dependência entre o termo, a classe e a entropia do termo. Ao aumentar com o valor da entropia do termo desvalorizar-se-ão os termos raros. Os termos raros apresentam uma entropia mais baixa do que termos mais frequentes, sendo que os termos raros poderão estar mais correlacionados com a classe do que os termos com maior entropia.

A medida *Gain Ratio* tem como objectivo normalizar IG, e assim, os temos raros não serão tão prejudicados por a sua baixa entropia. Logo, a diferença na expressão de GR e IG é que com GR a expressão encontra-se normalizada por um factor de multiplicação (Debole e Sebastiani s.d.).

2.2.9. *Mutual Information*

A técnica MI, *Mutual Information*, é muito utilizada em modelação de linguagem. MI visa realizar associações entre termos aleatoriamente escolhidos, e nesse processo determinar a dependência que esses termos têm entre si (Yiming e Jan s.d.).

MI é calculado por:

$$I(t,c) = \log \frac{P_r(t \cap c)}{P_r(t)P_r(c)} \tag{2.8}$$

Sendo t o termo e c a classe. A expressão (2.8) poderá ser traduzida no contexto de categorização de textos da seguinte forma:

$$I(t,c) \approx \log \frac{A.N}{(A+C)(A+B)} \tag{2.9}$$

A representa o número de vezes que a termo t e a classe c co-ocorrem;
B representa o número de vezes que o termo t ocorre sem ser na classe c;
C representa o número de vezes que a classe c ocorre sem o termo t; e
N representa o número total de documentos.

MI e IG têm uma complexidade temporal muito próximas (Yiming e Jan s.d.).
A desvantagem de MI deve-se ao facto dos seus pesos serem influenciados pelas percentagens marginais dos termos. Isto devido à seguinte característica de MI:

$$I(t,c) = \log P_r(t|c) - \log P_r(t) \tag{2.10}$$

Os termos com a mesma probabilidade condicional, $P_r(t|c)$, irão pontuar os termos raros com maior peso do que os termos comuns. Estes pesos não serão comparáveis com outros termos, os mais comuns, porque em oposição aos termos raros possuem uma frequência muito mais variável (Yiming e Jan s.d.).

2.2.10. *Term Strength*

TS é radicalmente diferente das técnicas, já mencionadas, IG, TF, CHI e MI. A pontuação obtida por TS depende do agrupamento[28] de documentos pois pretende encontrar o número de palavras comuns entre documentos, assumindo que quanto maior for o número de palavras comuns mais relacionados estarão os documentos.

[28] Do Inglês, *clustering*.

Para o cálculo de TS é necessário indicar o valor de *threshold*[29] pretendido, este valor irá indicar o número mínimo palavras comuns (similaridade) entre dois documentos. Assim se saberá o quão próximo os documentos se encontram. O tempo computacional gasto com TS é quadrático sendo dependente do número de documentos treino a utilizar.

Sendo d_1 e d_2 dois documentos relacionados aleatoriamente escolhidos, e t um termo, o TS é medido da seguinte forma (Yiming e Jan s.d.):

$$TS(t) = P(t \in d_1 | t \in d_2) \tag{2.11}$$

2.2.11. *GSS coeficiente*

GSS trata-se de uma simplificação do método CHI proposta por Galavotti (Zheng, Srihari e Srihari, A Feature Selection Framework for Text Filtering s.d.).
Com GSS os termos positivos correspondem a termos que pertencem à classe, e termos negativos indica que não pertencem. Sendo t um termo e c_i a classe i, GSS é determinado pela seguinte expressão:

$$GSS(t, c_i) = P(t, c_i)P(\bar{t}, \bar{c_i}) - P(t, \bar{c_i})P(\bar{t}, c_i) \tag{2.12}$$

Em que t representa a ocorrência do termo t, e $\bar{t}$ a não-ocorrência do termo t; e c representa a ocorrência da classe c, e $\bar{c}$ a não-ocorrência da classe c.

[29] *Threshold* representa um limiar. Apenas a partir desse valor será considerado, ou só será considerado até atingir esse valor.

2.2.12. Terceiro Momento em relação à média

Como foi referido no capítulo 1 o Terceiro Momento é uma medida que foi proposta recentemente, havendo, por isso, um interesse especial em estudá-lo e compará-lo com as restantes medidas de selecção de termos, utilizando diferentes representações de documentos.

$$V(t) = \frac{\frac{1}{C.\, p_m(t,.)} \sum_{c=1}^{c=C} [p_m(t, classe_c) - p_m(t,.)]^3}{Disp_m(t)} \tag{2.13}$$

C: Número total de classes .
$p_m(t, c)$: Probabilidade média de um termo ocorrer na classe c.
$p_m(t, .)$: Probabilidade média de um termo ocorrer em todas as classes do *corpus*.
$disp_m(t)$: Dispersão[30] média do termo nas classes.

Na equação 2.13 pode visualizar-se como se determina o Terceiro Momento. A probabilidade média de um termo t ocorrer na classe i, e em todas as classes do *corpus* é determinado da seguinte forma:

$$p_m(t, c) = \frac{1}{\#\text{docs da classe c}} \sum_{j=1}^{j=\#\text{docs classe c}} \frac{f(t, doc_j)}{\#\text{palavras } doc_j} \tag{2.14}$$

$f(t, doc_j)$: A frequência de um termo t no documento j.

$$p_m(t, .) = \frac{1}{\#classes} \sum_{i=1}^{i=\#classes} p_m(t, classes_i) \tag{2.15}$$

#classes: Número total de classes.

[30] Disperso no sentido de ter uma variação elevada.

A dispersão dos termos nas classes irá sobrevalorizar os termos que são mais fiéis às classes. Sendo a dispersão de um termo t determinado a partir das seguintes expressões:

$$\mathrm{disp}(t, \text{classe i}) = \frac{1}{\#\text{docs da classe i} * p_m(t,.)} \sum_{j=1}^{j=\#\text{docs classe i}} \left| \left[p_m(t, \mathrm{doc}_j) - p_m(t,.) \right]^3 \right| \tag{2.16}$$

$p_i(t, \mathrm{doc}_j)$: Probabilidade de um termo t ocorrer no documento j da classe i.
$p_i(t,.)$: Probabilidade de um termo t ocorrer em todos os documentos da classe i.

$$\mathrm{Disp}_m(t) = \frac{1}{\#\text{classes}} \sum_{i=1}^{i=\#\text{classes}} \mathrm{disp}(t, \mathrm{classe}_i) \tag{2.17}$$

2.3. Agrupamento vs Classificação

A fase de teste de um classificador, tem a função de comparar a representação do documento a classificar (muitas vezes descrita num vector) com os padrões aprendidos na fase de treino e, a partir disso, atribuir-lhe uma das classes.

O processo de agrupamento é similar à fase de treino dos classificadores. No entanto, o processo de agrupamento não poderá ser Supervisionado pois iria contra a sua natureza**.** Agrupamento, apenas, poderá através de diferentes técnicas observar padrões. Sem ser indicado a qual classe os documentos pertencem, esta descoberta baseia-se unicamente na observação do conteúdo dos documentos.

Concluindo, podemos dizer que a Classificação atribui documentos a classes (tópicos) pré-definidos, enquanto o agrupamento consiste normalmente em atribuir documentos a classes conforme a distribuição dos restantes documentos de treino.

2.4. Algoritmos de Agrupamento

Independente da natureza dos elementos a agrupar, existem várias técnicas de agrupamento, às quais serão referidas brevemente.

No caso específico do agrupamento de documentos, a organização dos grupos é feita com base nas semelhanças e diferenças entre os documentos. Pretende-se que os documentos de um mesmo grupo partilhem características comuns. A cada grupo associa-se normalmente uma classe. Estas classes poderão identificar padrões, comportamentos, tópicos, línguas, entre muitos outros. No contexto desta dissertação, um grupo corresponde a um tópico.

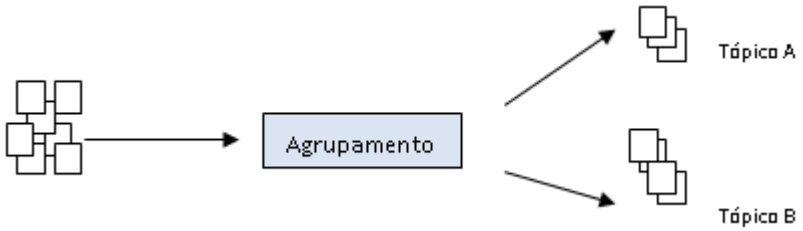

Figura 2.4. 1: Agrupamento

2.4.1. Agrupamento hierárquico

Este método de agrupamento realiza a descoberta da hierarquia dos agrupamentos (*clusters*) sem o auxílio do utilizador, sendo esta uma vantagem perante técnicas não-hierárquicas. Técnicas não-hierárquicas normalmente exigem que o utilizador pré-defina o número de classes – clusters –. Contudo, as técnicas não-hierárquicas tendem a ser menos pesadas computacionalmente.

Os métodos hierárquicos dividem-se em duas categorias: aglomerativos ou divisivos. Os métodos hierárquicos aglomerativos visam agrupar os documentos em *clusters* (agrupamentos) cada vez maiores, ou seja, agrupam até que todos os dados façam parte do mesmo grupo. Os métodos divisivos realizam a tarefa oposta. Estes inicialmente contêm o conjunto de documentos num só agrupamento e, a partir deste, subdividem-no em diferentes agrupamentos até que cada agrupamento contenha um só elemento.

Os métodos divisivos podem ser monotéticos ou politéticos. Aquando da decisão de divisão do agrupamento os métodos monotéticos apenas consideram uma dimensão de representação, funcionando como variáveis binárias perante presença/não-presença dessa dimensão num dado objecto (documento). No caso dos métodos politéticos já é tido em conta um conjunto de dimensões, todas elas relevantes são pesadas no momento de divisão do agrupamento.

Os métodos aglomerativos apresentam diferentes métodos de decisão de agrupamento, sendo os mais conhecidos o *single-link*, o *complete-link*, o *average-link*, entre outros. O que varia entre estes métodos é o cálculo da proximidade entre os documentos a agrupar (Muscat s.d.):

1. *Single-link* – A distância mínima entre dois clusters é a distância mínima entre quaisquer dois elementos de cada cluster.

2. *Complete-link* – A distância entre dois clusters é a distância máxima entre dois elementos dos dois clusters.

3. *Average-link* – A distância entre dois clusters é a média das distâncias entre todos os elementos de cada um dos clusters.

2.4.3. K-*means*

Trata-se de um método de partição que recebe os dados totais e o número de clusters a formar. São determinados k centróides aleatórios, sendo k o número de clusters indicado previamente, a que cada elemento do conjunto de dados é associado ao centróide mais próximo - distância euclidiana -. Cada vez que um elemento é adicionado a um cluster é necessário recalcular o novo centróide do grupo. Este processo repete-se até se encontrar uma situação de paragem (Sambasivam e Theodosopoulos s.d.).

Este método apresenta duas grandes limitações, independentemente do âmbito da aplicação. Em primeiro lugar, exige que o utilizador indique o número de clusters (grupos), o que se torna uma limitação no caso das abordagens em que não se tem conhecimento prévio do número de grupos existentes no *corpus*. Em segundo lugar, e uma vez que este método utiliza a distância euclidiana para atribuir a classe ao novo documento na fase de classificação, este método assume que os grupos tendem a formar-se no espaço a várias dimensões segundo nuvens hiper-esféricas de igual tamanho. Esta assumpção só coincide com a realidade em casos raros, o que torna este método relativamente pouco preciso.

2.4.4. K-*medoids*

Semelhante ao K-*means* mas, em vez de usar um centróide calculado pela média dos elementos, usa um dos elementos como representante do grupo. No entanto é possível escolher o critério para determinar qual a distância entre elementos.

2.5. Algoritmos de Classificação

Genericamente um algoritmo de classificação tem como objectivo separar os dados de entrada em classes distintas. No entanto, o meio como se atinge este objectivo nem sempre é o mesmo, e por vezes quando se comparam dois algoritmos observa-se que são muito distintos entre si.
Brevemente serão descritos alguns algoritmos, e facilmente se poderá observar as diferenças mais óbvias entre estes.

2.5.1. Classificadores Probabilísticos

Denomina-se por classificador probabilístico o classificador que utiliza probabilidades para decidir se um objecto pertence a uma determinada classe.
Assim, um classificador probabilístico estimará para cada classe C a probabilidade de um objecto lhe pertencer, em que $C \in \{C_1, \dots, C_{|C_{TotalClasses}|}\}$. Estes objectos encontram-se representados por um vector dos termos característicos do seu conteúdo. No âmbito deste trabalho o objecto será o documento, e as classes serão os tópicos.

Nos classificadores probabilísticos é muito usada a técnica de representar os objectos por pesos binários: 0 indica que um termo não se encontra num documento e 1 indica que o documento contém o termo (Sparck e Robertson s.d.).

Naïve Bayes é um classificador muito conhecido e já aplicado em várias abordagens (Teevan e Jason s.d.), (Rish, Hellerstein e Thathachar s.d.), (Pop s.d.) entre outras. Este classificador assume a independência entre os termos de um documento. Apesar desta hipótese de independência ser uma hipótese que simplifica a classificação, é por vezes muito criticada por não reflectir a realidade.

Os classificadores probabilísticos, sendo *Naïve* Bayes apenas um dos exemplos, tendem a ser criticados por realizarem uma análise puramente quantitativa. Esta análise não facilita uma futura

interpretação feita pelas pessoas. Outros algoritmos que se baseiam em árvores e em regras de decisão, que a seguir serão descritos, já não sofrem deste tipo de críticas.

2.5.2. Árvores de decisão

Um classificador em árvore, ou hierárquico, é um algoritmo, que perante um problema, o subdivide em diversos problemas mais simples. Os níveis em que as diferenças são mais óbvias serão os primeiros a ser identificados. Nos níveis inferiores serão detectadas as diferenças mais subtis.

Uma árvore de decisão contém os termos nos nós internos. As folhas indicam o tópico, e os ramos representam os valores possíveis que o objecto (documento) pode tomar por cada termo. Os documentos são testados pelos valores possíveis, presença ou ausência dos termos (nós), e perante essa decisão são direccionados para outro ramo da árvore. Aquando da chegada à folha (tópico) não existe desdobramento: é-lhe atribuída a classe da folha ao documento (Abreu s.d.).

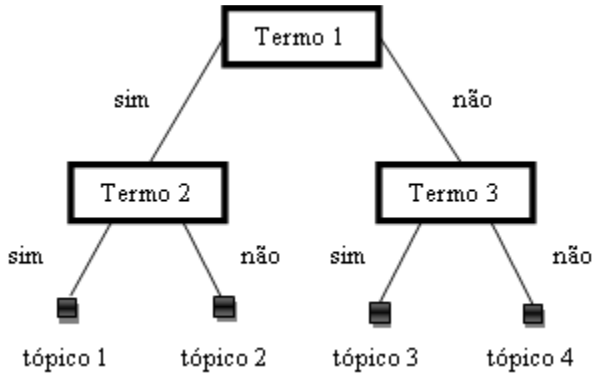

Figura 2.5.2. 1: Exemplo de uma árvore de decisão.

A maior dificuldade neste algoritmo é a decisão de repartição. A escolha dos termos que melhor divide o conjunto em subconjuntos é essencial. É importante que os termos sejam decisivos, ou seja, pertence ou não-pertence. Obter uma solução para este problema é muito difícil. O número

de estruturas de árvores possíveis, ou seja as diferentes partições do conjunto inicial para cada classificação em diferentes classes, é muito elevado.

2.5.3. Regras de decisão

Uma alternativa ao algoritmo de árvores de decisão será a de utilizar regras de decisão. Estas regras são muito similares às árvores pois podem aplicar qualquer função booleana. No entanto este algoritmo permite gerar regras muito mais compactas (Apté, Damerau e Weiss s.d.).

2.5.4. Redes neuronais

Uma das razões que incentivaram o aparecimento de redes neuronais artificiais foi o de se pretender simular o funcionamento do cérebro humano. De acordo com o conhecimento científico disponível, o cérebro humano consiste numa rede de unidades elementares denominadas neurónios, intensamente interligados. Esta abordagem baseia-se num modelo matemático do funcionamento do neurónio, sendo uma rede neuronal artificial um conjunto destes elementos (nós) interligados.

Uma possibilidade de implementação de uma rede deste tipo consiste num conjunto de nós que se encontra dividido em camadas, existindo uma camada de entrada, um conjunto de camadas denominadas escondidas, e uma camada de saída. A cada nó estão associados parâmetros (pesos) relativos a cada outro nó a si ligado, e são esses parâmetros que são ajustados durante o chamado processo de treino. Num caso de classificação, por exemplo, o sistema é treinado com um conjunto de dados e respectivas classificações. Durante o processo de treino os pesos vão sendo ajustados de modo a minimizar a diferença entre a saída efectiva da rede e a saída que ela deveria produzir. Uma vez treinada, a rede deveria ser capaz de classificar correctamente exemplos semelhantes (Fonseca s.d.).

Considerando o caso dos classificadores, o início do processo de treino pressupõe um conjunto de pesos iniciais, que em geral é obtido de um modo aleatório. Por esse motivo um mesmo conjunto de treino pode produzir classificadores distintos relativamente ao seu desempenho.
Estes sistemas confrontam-se com limitações computacionais, visto que muitos problemas exigem elevados desempenhos nesse domínio. Esta abordagem apresenta bons resultados em diversos domínios e é expectável que venha a tornar-se cada vez mais útil à medida que as capacidades computacionais forem evoluindo.

2.5.5. Método de Rocchio

O objectivo do método de Rocchio será representar um perfil/classe (tópicos) através de um vector próprio e a partir de um conjunto de dados de treino (Sebastiani, Machine Learning in Automated Text Categorization s.d.).

Inicialmente, o algoritmo aprende, através de exemplos positivos e negativos, qual o vector próprio (característico) de cada classe. A determinação da classe de um documento é feita com base na comparação entre o vector característico de cada classe com o vector do documento.
Os vectores próprios das classes $\overrightarrow{c_i} = \langle w_{1i}, \dots, w_{|\tau|i} \rangle$, são determinados a partir da média de todos os vectores dos documentos de treino para uma determinada classe:

$$w_{ki} = \beta * \sum_{\{d_j \epsilon\, POS_i\}} \frac{w_{kj}}{|POS_i|} - \gamma * \sum_{\{d_j \epsilon\, NEG_i\}} \frac{w_{kj}}{|NEG_i|} \tag{2.18}$$

Sendo:
w_{ki} - o peso do termo t_k no vector próprio i.
w_{kj} – o peso do termo t_k no docomento j.
POS – o número de documentos de teste positivos.
NEG – o número de documentos de teste negativos.
β e γ - são parâmetros de controle que permitem dar importâncias relativas aos exemplos positivos e negativos.

A maior desvantagem deste classificador será a de identificar uma classe com um centróide. Caso estas classes apresentem uma forma mais dispersa, os resultados não serão bons visto o centróide se encontrar mais afastado dos documentos.

2.5.6. Classificadores baseados em exemplos

Um classificador que seja baseado em exemplos é um classificador que não identifica uma representação da classe. Depende da observação de características similares entre os documentos de treino. Estes métodos também são conhecidos como *lazy learners* (Wikipedia s.d.).

2.5.7. Método dos vizinhos mais próximos

O algoritmo dos k-vizinhos mais próximos é um processo de aprendizagem baseado em instâncias englobando-se nos métodos de classificadores baseados em exemplos.
Para decidir se um documento d deve ser classificado na classe c o algoritmo verifica se os *k* documentos de treino mais próximos pertencem à classe c. Produz melhores resultados quando tem uma grande quantidade de dados de treino, não tendo, no entanto a capacidade de identificar atributos (termos) que possam ser irrelevantes para a classificação (Ozgur n.d.).

2.5.8. Máquinas de Vectores de Suporte

Os algoritmos de Máquinas de Vectores de Suporte[31] apareceram como uma técnica para auxiliar a identificação de padrões num conjunto de dados.

Nesta técnica os dados de entrada (documentos de treino) são transformados em vectores de termos. Estes vectores serão projectados linearmente num espaço de |N|-dimensões, em que N - número de dimensões do espaço - poderá variar conforme o número de termos a serem tomadas em consideração.

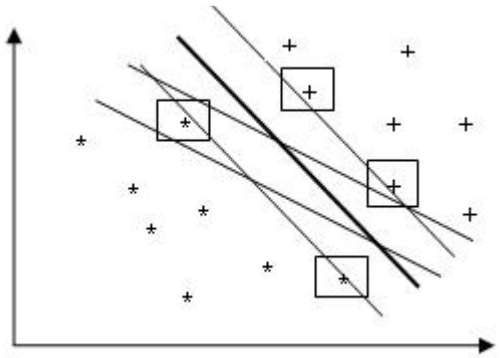

Figura 2.5.8. 1: Exemplo ilustrativo de uma *Support Vector Machine*.

O objectivo deste algoritmo será separar linearmente os termos (características) através de superfícies de decisão[32], em que os termos que se encontrem mais próximos do limite entre estas superfícies serão seleccionados; os Vectores de Suporte realçados na figura 2.5.8.1. São aqueles que definem a margem de separação óptima entre as classes. (Baranoski s.d.).

[31] Do inglês, *Support Vector Machine*.
[32] Do inglês, *decision surfaces*.

2.5.9. RIPPER

RIPPER trata-se de uma variante dos algoritmos de classificação que utilizam regras de decisão (Capítulo 2.5.3).

Cohen e Singer (Cohen e Singer s.d.) indicam que se trata de um algoritmo eficiente no tratamento de um elevado volume de *noisy data*[33], ou seja, documentos que não contêm qualquer relevância para a classificação. Processa os dados em tempo linear, ou muito próximo do linear.

RIPPER representa os documentos como uma lista ordenada de *tokens*[34], não sendo necessário extrair do corpus um subconjunto com os termos/dimensões mais relevantes. Esta técnica de traduzir directamente os documentos em *tokens* é denominada por *direct representation*.

Com RIPPER a presença ou ausência de uma palavra influencia a classificação, ou seja, a ausência de uma palavra não é ignorada. Conforme o contexto de uma palavra a sua ausência irá pesar na classificação.

Este classificador determina o contexto de uma palavra por uma conjunção realizada da seguinte forma:

$$w_1 \in document\ and\ w_2 \in document \ldots and\ w_k \in document \tag{2.19}$$

Assim, pode-se afirmar que com RIPPER o contexto de uma palavra w_1 consiste em um conjunto de outras palavras w_2 … w_k que têm de co-ocorrer com w_1. Podendo ocorrer em qualquer ordem e em qualquer localização no documento.

[33] *Noisy Data* neste contexto entende-se como um conjunto de dados que contêm muita informação não-relevante para os objectivos presentes. Este tipo de dados poderá dificultar a captação de informação relevante, e consequentemente a performance do algoritmo.

[34] *Token* em computação é um segmento de texto ou símbolo que pode ser manipulado por um *parser*, que fornece um significado ao texto; em outras palavras, é um conjunto de caracteres com um significado colectivo (Wikipedia s.d.).

$$irlanda \leftarrow irlanda \in documento$$
$$irlanda \leftarrow ira \in documento, morto \in documento$$
$$irlanda \leftarrow ira \in documento, mortes \in documento$$
$$irlanda \leftarrow ira \in documento, belfast \in documento$$
$$irlanda \leftarrow irland\hat{e}s \in documento, aborto \in documento$$
$$irlanda \leftarrow ira \in documento, tiro \in documento$$
$$else\ not\_irlanda$$

Fig. 2.5.9.1: Um conjunto de regras de decisão para a categoria 'Irlanda'

RIPPER trata-se de um classificador composto por um conjunto de regras, como já foi referido acima. Na figura 2.5.9.1 encontra-se um exemplo de como este classificador interpreta a informação dos documentos através de regras de decisão (Cohen e Singer s.d.).

2.5.10. Classificador Proposto por Joaquim F. Silva

No trabalho do Professor Joaquim F. Silva é indicada a técnica do Terceiro Momento em relação à média, e um classificador construído para esta medida. No classificador proposto, após o cálculo de V(t) (Secção 2.2.12) para cada termo é utilizada a seguinte abordagem:

- Definição de uma matriz de semelhanças entre documentos. A ideia reside na possibilidade de cada documento ser caracterizado pelos seus pares, isto é, pelos outros documentos.

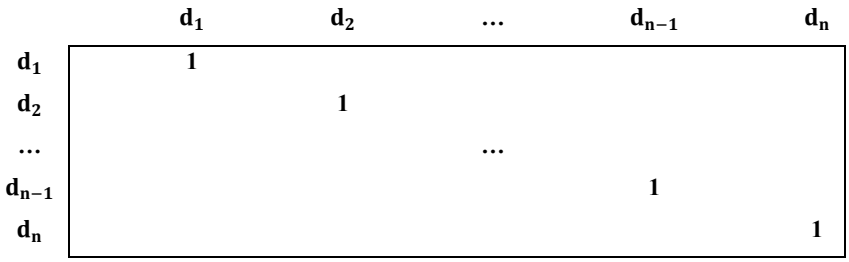

Figura 2.5.10. 1: Matriz de Co-variâncias

n: Número total de documentos de treino .

Assim, para cada documento i, será determinada a sua semelhança com os restantes documentos. Por exemplo, a semelhança entre o documento i e o documento j será determinada da seguinte forma:

$$Sim(d_i, d_j) = \frac{cov\ (d_i, d_j)}{\sqrt{cov\ (d_i, d_i)} * \sqrt{cov\ (d_j, d_j)}} \tag{2.20}$$

Em que:

$$cov(d_i, d_j) = \frac{1}{|T|} \sum_{t \in T} \left[\left(p^*(t, d_i) - p^*(., d_i) \right) * \left(p^*(t, d_j) - p^*(., d_j) \right) \right] \tag{2.21}$$

Sendo:

$$p^*(t, d_i) = p(t, d_i) * V(t) \tag{2.22}$$

$$p(_{.,d_i}) = \frac{1}{|T|} \sum_{t \in T} p(t, d_i) \tag{2.23}$$

t: Termo/Dimensão.

V(s): Valor discriminante da sequência (termo) s.

$P(t, d_i)$: Indica a probabilidade de um termo ocorrer no documento i.

$p^*(t, d_i)$: Probabilidade de um termo ocorrer no documento i tendo em conta o valor discriminante do termo t que fora determinado anteriormente.

$p(., d_i)$: Probabilidade média de ocorrências de todos termos no documento i.

Utilizam-se frequências relativas (probabilidades de ocorrência) de termos em cada documento ao invés da sua frequência absoluta, de forma a que o tamanho deste não influa na importância relativa do termo no documento.

Na matriz os valores irão variar entre -1 e 1, sendo que:

a. O valor de 0 indica que os documentos não contêm nenhuma semelhança entre si;
b. O valor de 1 só deverá ocorrer quando se trata do mesmo documento;
c. Valores negativos indicam um valor de dissimilhança[35] entre os documentos;
d. Valores positivos indicam quão semelhantes são os documentos entre si.

Espera-se visualizar valores positivos e próximos de 1 para documentos da mesma classe e valores menores entre documentos de classes diferentes.

[35] Entende-se por dissimilhança o quão diferentes dois documentos possam ser entre si.

2.6. Ferramenta WEKA

Weka[36] (*Waikato Environment for Knowledge Analysis*) trata-se de uma ferramenta muito utilizada em *Machine Learning* e *Data Mining*. Weka é um software disponível em *open-source* (Wikipedia s.d.), e implementada em ambiente Java.

Weka (Witten, et al. s.d.) disponibiliza muitas abordagens de *Machine Learning*. Os seus fortes encontram-se na área da classificação, onde muitos investigadores (Hall e Holmes s.d.) realizaram o seu trabalho com o seu auxílio. Weka além de disponibilizar vários classificadores como Naive Bayes, K-Nearest Neighbour, RIPPER, entre muitos outros, ainda possibilita: definir o (s) ficheiro (s) de treino; determinar a precisão e o recall do algoritmo; gerar a matriz de confusão; entre outros.

Entre os formatos admitidos por a ferramenta Weka encontra-se o ARFF. Trata-se de um ficheiro de texto em formato ASCII em que é enumerado uma lista de instâncias que partilham atributos. No contexto desta dissertação as instâncias irão corresponder a documentos e os atributos a termos e classes.

```
% Faculdade de Ciências e Tecnologia de Lisboa
% Departamento de Informática
% Orientador: Professor Doutor José Gabriel P. Lopes
% Autor: Filipa Peleja
% Data: Julho 2009

@RELATION Palavras

@ATTRIBUTE palavra1 NUMERIC
@ATTRIBUTE palavra2 NUMERIC
                                …
@ATTRIBUTE class {classe1, classe2, …, classe92}

@DATA
2,0,4,1,0,0, …, 2, classe1
1,0,0,2,1,0, …, 0, classe1
0,2,1,8,0,0, …, 2, classe10
                                …
```

Figura 2.6. 1: Exemplo de um ficheiro em formato ARFF

[36] Weka encontra-se disponível em http://www.cs.waikato.ac.nz/ml/weka/.

Na figura 2.6.1. é possível visualizar como é estruturado um ficheiro em formato ARFF. Cada atributo, ou termo, estará associado a um valor numérico com excepção do atributo class. Para os termos será contabilizado o número de vezes que ocorre em cada documento estando, por isso, associado a um valor numérico.

A classe corresponderá à palavra que a representa. Se no *corpus* os documentos pertencessem a um universo de três classes: História, Ciência e Línguas, o atributo class seria representado por ”@ATTRIBUTE class {historia, ciencia, linguas}”.

Nas linhas que seguem à instrução @DATA será inserida a informação referente a cada documento. Cada posição da linha corresponde a um inteiro que representa o número de ocorrências de cada termo no documento, tendo por fim a classe a que pertence.

2.7. Conclusões obtidas em trabalho realizado por outros autores

Yang e Pedersen (Yiming e Jan s.d.) realizaram um estudo comparativo da performance de várias técnicas de selecção de termos com o objectivo de conseguirem avaliar o seu desempenho sobre as mesmas condições, sendo o conjunto de documentos a categorizar de grande dimensionalidade. Uma das razões que motivou estes autores a realizarem este trabalho foi porque, até a esse momento, muitas das técnicas de selecção de termos apenas tinham sido testadas sobre corpus de dados de baixa dimensionalidade, e por isso não existiam garantias de uma boa performance sobre situações mais exigentes.

Yang and Pedersen descobriram fortes correlações entre as técnicas TF(*Term Frequency*), IG (*Information Gain*) e CHI (*Chi Square*). Esta aproximação entre estas técnicas encontra-se no valor que atribuem aos termos, sendo estes pontuados com valores muito próximos.

Também observaram que as técnicas TF, IG e CHI foram as que demonstraram melhor performance. Devido a isso concluíram que os termos comuns entre os documentos são de facto informativos e importantes para a categorização de documentos. São técnicas que têm em consideração a distribuição dos termos pelos diferentes documentos.

A aproximação dos resultados de TF, IG e CHI sugerem que TF, apesar da sua simplicidade poderá ser fiável para uso ao invés de técnicas mais complexas, e mais exigentes computacionalmente. Também observaram que TS (*Term Strength*) é uma técnica competitiva até atingir 50% da redução da dimensionalidade dos termos. TS necessita de um número maior de termos para conseguir apresentar uma performance próxima de técnicas como IG, TF e CHI.

Confirmaram que MI (*Mutual Information*) teve um desempenho muito baixo, e indicaram que o motivo porque isso se passou foi devido ao facto de MI favorecer os termos raros. E também devido à sua sensibilidade a erros derivados de estimar probabilidades.

No trabalho de Pekar, Krkoska e Staab (Krkoska, Pekar e Staab s.d.) foram exploradas várias hipóteses de optimização de diversas técnicas de selecção de termos: MI, GR (*Gain Ratio*), OR (*Odds Ratio*) e TSL_1 e TSL_2 duas variantes de TS.

Com TSL_1 as palavras relacionadas, com um termo a pesar, são pesquisadas apenas nos documentos da classe e não no conjunto de todos os documentos de treino, e apenas considera os termos mais próximos. Assim, determina o peso do termo por estimar a sua distribuição nos pares mais próximos da seguinte forma:

$$TSL_1(t,c) = P(t \in n | t \in n'), \qquad com\ \ n, n' \in c \tag{2.24}$$

Sendo que t representa o termo a pesar, e c a classe. Com o auxílio de uma medida de similaridade foi determinado um conjunto de pares de palavras (n,n'), em que estes pares representam as palavras que se encontram relacionados. Esta relação irá ajudar na determinação do peso a se atribuir a t. TSL_1 visa aumentar a similaridade entre membros da mesma classe e descartar similaridades entre classes.

TSL_2 tem como objectivo superar o problema do elevado custo computacional de TSL_1. A comparação com apenas os vizinhos mais próximos de um termo a pesar causa uma perca muito elevada de processamento do computador. TSL_2 ao invés de utilizar apenas os vizinhos mais próximos do termo utiliza todas os termos que pertencem à classe.

$$TSL_2(t,c) = \frac{|\{n \in c | t \in n\}|}{|\{n \in c\}|} \tag{2.25}$$

No estudo de (Krkoska, Pekar e Staab s.d.) foram comparados resultados obtidos com as várias técnicas de selecção de termos; utilizando o valor máximo local da relevância de um termo por todas as classes[37]; o valor médio da sua contribuição por cada classe (tendo em consideração o tamanho da classe); e a soma de todos os valores locais para cada termo.

[37] Neste trabalho foram utilizados os *corpus* de *British National Corpus* (British National Corpus s.d.) e *Associated Press 1988 Corpus* (Datasets from Some Distributional Similarity Experiments s.d.).

Os resultados indicaram uma melhor performance quando se utilizava o valor máximo local. Usualmente um termo tem valores mais elevados com algumas classes em específico, enquanto nas restantes o seu peso é muito baixo. Se for utilizada uma média do seu peso sobre todas as classes, ou a soma de todos os pesos, a sua relevância sobre as classes que mais se identifica será retirada. O valor máximo local realça o peso dos termos nas classes com que mais se identificam. As técnicas MI, GR e OR demonstraram melhores resultados aquando da utilização de variáveis globais, ou seja, não sendo específicas (locais) da classe em relação às técnicas TSL_1 e TSL_2. Os autores indicam que o motivo porque isso se passou deveu-se ao facto dos termos utilizados por estas técnicas, que têm maior peso, serem muito discriminantes para a classe; muitas vezes são raras ocorrendo em apenas uma classe e não causando atrito entre termos de classes diferentes que foram pesadas com o mesmo peso. É também vantajoso pesar os termos globalmente porque assim garante que a maior parte dos termos têm um valor superior a zero (aproximam-se de, pelo menos, uma classe, por muito baixo valor que seja).

TSL_1 e TSL_2 são mais vantajosas em situações em que se pretenda distinguir com maior facilidade uma classe de outra. Estas técnicas por retornarem melhores resultados aquando da utilização de termos locais conseguem com maior facilidade discriminar a diferença entre classes. Os autores (Krkoska, Pekar e Staab s.d.) afirmam que apesar dos termos individualmente não serem suficientes para diferenciar as classes, quando se tem um conjunto já serão muito úteis.

No trabalho realizado por Debole e Sebastiani (Debole e Sebastiani s.d.) indicam que a técnica TFIDF não é a melhor opção para o seu trabalho. Os autores consideram que a informação obtida pela contribuição positiva e negativa é muito importante para os resultados finais, e, como já foi referido anteriormente, a técnica TFIDF descarta as contribuições negativas.

Debole e Sebastiani (Debole e Sebastiani s.d.) observaram que para classificadores SVM as técnicas GR e CHI são as que demonstram melhor desempenho, demonstrando a sua superioridade de 11% sob TFIDF. Consideraram ainda que, apesar de IG ser muito próximo de GR, os seus resultados foram muito desapontadores quando aplicado globalmente. Quando é utilizado localmente os resultados são muito melhores e, por vezes, até superando a técnica CHI.

2.8. Medidas para avaliar classificadores

Realizar a avaliação dos classificadores pode ser uma tarefa complicada. Supondo que se fez uma classificação sobre um conjunto de documentos sobre o qual nada se sabia, será difícil avaliar, de forma precisa, o trabalho do classificador.

No entanto, pode-se utilizar um conjunto de dados conhecidos (conjunto de teste), avaliando-se assim o desempenho do classificador. Para tal são normalmente utilizadas medidas como: *Precision*; *Recall*; *F-Measure*; Matriz de Confusão; Exactidão; Estatística *Kappa*; *Relative Operating Characteristic* (ROC); e *Micro-Averaging*. Nas próximas secções descrevem-se sucintamente estas medidas.

2.8.1. *Precision* e *Recall*

Para definir estas medidas é necessário introduzir alguns conceitos acerca dos resultados de uma classificação. Relativamente a uma determinada classe A, considera-se um Verdadeiro Positivo a classificação na classe A de um documento que de facto pertence a essa classe. VP representa o número de Verdadeiros Positivos numa classificação. Um Verdadeiro Negativo representa o caso em que um documento que não pertence à classe A é classificado como não pertencente a essa classe. VN representa o número de Verdadeiros Negativos numa classificação. A classificação de um documento como pertencente à classe A, mas que de facto não pertence a essa classe, denomina-se um Falso Positivo. FP representa o número de Falsos Positivos numa classificação. Por fim, a classificação de um documento como não pertencente à classe A, mas que de facto lhe pertence, denomina-se Falso Negativo. FN representa o número de Falsos Negativos numa classificação.

Precision é dada pela expressão,

$$precision = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FP}} \tag{2.26}$$

Esta medida é um indicador da exactidão do classificador, no sentido em que o valor máximo de 1, significa que todos documentos classificados como pertencentes a uma determinada classe pertencem efectivamente a essa classe, ou seja, não há Falsos Positivos. No entanto, nada informa acerca dos documentos desta classe que tenham eventualmente sido erradamente classificados como não lhe pertencendo (Falsos Negativos).

Recall, que é dado pela expressão,

$$recall = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FN}} \tag{2.27}$$

é uma medida de plenitude, no sentido em que o valor máximo de 1, significa que todos documentos que pertencem a uma determinada classe foram classificados como pertencentes a essa classe, ou seja, não há Falsos Negativos. Mas não informa sobre os documentos, que não pertencendo à classe, foram incorrectamente classificados como pertencendo (Falsos Positivos).
Estas duas medidas dão informação complementar, pelo que normalmente são utilizadas em conjunto na avaliação de classificadores.

2.8.2. F-Measure

Como se viu, as duas medidas descritas na secção anterior dão informação complementar sobre o desempenho do classificador. *F-Measure* é a média harmónica de *Precision* (P) e *Recall* (R), ou seja,

$$F - Measure = \frac{2.precision.recall}{precision + recall} = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} \tag{2.28}$$

combinando assim a informação das duas medidas.

A *precision* apenas tem em consideração os documentos que foram correctamente classificados, por isso, acarreta alguns custos. O cálculo desta medida não é afectado por os documentos que pertencentes a uma classe foram incorrectamente classificados como pertencentes a outra classe. A medida *recall* não sofre do mesmo problema, porém, os documentos que foram classificados como pertencentes à classe, e na realidade pertencem a outra classe, não são tidos em consideração.

F-Measure apenas apresenta valores elevados quando *precision* e *recall* apresentam valores elevados. Por isso, esta medida tem em consideração os documentos, que incorrectamente foram classificados como pertencentes à classe, e que incorrectamente foram classificados como pertencentes a outra classe.

2.8.3. Matriz de confusão

A matriz de confusão é um instrumento fundamental na análise do desempenho de classificadores. Trata-se de uma matriz quadrada de dimensão NxN, em que N é o número de classes, que disponibiliza informação por classe sobre o número de documentos correctamente e incorrectamente classificados.

	Classificados como			
	Classe 1	Classe 2	…	Classe N
Classe 1	k	w	…	j
Classe 2	h	z	…	i
…	…	…	…	…
Classe N	l	y	…	m

Tabela 2.8.3. 1: Matriz de Confusão para N classes

Cada posição de uma linha corresponde à classificação correcta, ou incorrecta, por cada uma das classes. Os valores indicados na diagonal representam os documentos correctamente classificados. Por exemplo, na tabela 2.8.3.1. o valor k corresponde aos documentos que foram classificados como pertencentes à classe da coluna 1 (Classe 1) e sendo na realidade pertencente à classe da linha 1, que é pois, a classe 1.

As posições de uma linha, com excepção da diagonal correspondem aos documentos incorrectamente classificados como não pertencentes à classe da linha, quando na realidade pertencem.

Outra visão sobre a Matriz de confusão será uma análise por coluna. Todas as posições da coluna, com excepção da posição pertencente à diagonal da matriz, correspondem a documentos que pertencem à classe da linha, mas que foram classificados como documentos pertencentes à classe da coluna. Por exemplo, na tabela 2.8.3.1. o valor h corresponde ao número de documentos classificados como da Classe 1, classe da coluna, mas que na realidade esses documentos pertencem à Classe2, classe da linha.

2.8.4. Exactidão

É uma medida de desempenho global do classificador, e representa a fracção de documentos que foram correctamente classificados, ou seja, corresponde ao somatório dos valores da diagonal da matriz de confusão dividido pelo número total de documentos classificados.

2.8.5. Estatística Kappa

A estatística Kappa (k), proposta por Cohen em 1960 (J. Cohen s.d.), e comprovada a sua aplicabilidade na área da classificação (Carletta s.d.), consiste numa medida estatística do grau de concordância entre classificações. Esta medida é calculada através da equação,

$$k = \frac{P_r(a) - P_r(e)}{1 - P_r(e)} \tag{2.29}$$

$P_r(a)$: Representa a probabilidade relativa da concordância entre as medidas utilizadas.

$P_r(e)$: Representa a probabilidade relativa de uma hipótese de concordância obtida aleatoriamente.

No caso em estudo interessa medir o grau de concordância entre a classificação efectuada por um classificador e a classificação efectiva do conjunto de documentos em análise. Na prática este cálculo é feito a partir da matriz de confusão que é construída no processo de classificação. Considerando n classes, e a matriz de confusão M(nxn), Pr(a) representa a probabilidade de um documento pertencente a uma determinada classe ser efectivamente classificado nessa classe. Esta probabilidade pode ser aproximada pelo número de classificações correctas dividido pelo número total de documentos, ou seja, a soma dos valores da diagonal da matriz de confusão dividida pelo número total de documentos, que é dada por,

$$P_r(a) = \frac{\sum_{i=1}^{n} M(i,i)}{\sum_{i=1}^{n}\sum_{j=1}^{n} M(i,j)} \tag{2.30}$$

M(i, i): Corresponde ao valor na posição da linha i e coluna i da matriz M.

M(i, j): Corresponde ao valor na posição da linha i e coluna j da matriz M.

Relativamente ao cálculo de Pr(e), note-se que dada a classe A, por exemplo, a probabilidade de um documento pertencer a essa classe, pode ser aproximada pelo somatório dos valores da linha da matriz de confusão referente a essa classe, dividido pelo número total de documentos. Seguindo o mesmo raciocínio, a probabilidade de um documento ser classificado na classe A, pode ser aproximada pelo somatório dos valores da coluna da matriz de confusão referente a essa classe, dividido pelo número total de documentos. Portanto, a probabilidade de um documento pertencer à classe A e ser classificado na classe A é o produto das duas probabilidades descritas acima. Pr(e) é a probabilidade de um documento pertencer a uma classe e ser classificado nessa classe, e pode assim ser aproximada pela expressão,

$$P_r(e) = \frac{\sum_{i=1}^{n} M(i,.)}{\sum_{i=1}^{n}\sum_{j=1}^{n} M(i,j)} * \frac{\sum_{j=1}^{n} M(.,i)}{\sum_{i=1}^{n}\sum_{j=1}^{n} M(i,j)} \tag{2.31}$$

$\sum_{i=1}^{n} M(i,.)$: Corresponde à soma de todos os valores da linha i da matriz M.

$\sum_{i=1}^{n} M(.,i)$: Corresponde à soma de todos os valores da coluna i da matriz M.

Em que o primeiro factor do lado direito é um vector linha constituído pelas somas das colunas divididas pelo número total de documentos, e o segundo factor é um vector coluna cujos elementos são as somas das linhas divididas pelo número total de documentos.

O valor de K varia entre -1 e 1: quão mais próximo estiver de 1 maior será a concordância, valores mais baixos indicam que a concordância que possa ter ocorrido foi devido a valores obtidos aleatoriamente, e por isso, não serão fiáveis. Um valor de 1 representa a concordância total, um valor de 0 caracteriza uma classificação aleatória, e um valor -1 representa a discordância total.

Valor de Kappa	Concordância
< 0	Não existe concordância
0 – 0.20	Ligeira
0.21 – 040	Considerável
0.41 – 0.60	Moderada
0.61 – 0.80	Substancial
0.81 – 1	Excelente

Tabela 2.8.5. 1: Valores de K com a medida Estatística Kappa

Ao contrário do cálculo simples das percentagens de concordância (Exactidão), este método tem em conta a concordância casual, pelo que em geral é considerado mais robusto. Existem no entanto autores (Strijbos, et al. s.d.) que consideram esta medida de concordância demasiado conservadora.

2.8.6. Relative Operating Characteristic (ROC)

Uma análise muito utilizada para medir o desempenho de um classificador é a análise da característica operacional do receptor - curva ROC (Braga s.d.). A curva ROC relaciona a taxa dos verdadeiros positivos V_P, no eixo dos yy', com a taxa de falsos positivos F_P, no eixo dos xx'. No contexto desta dissertação, verdadeiros positivos correspondem aos documentos que foram correctamente classificados, e falsos positivos correspondem aos documentos que foram classificados como se fossem de uma classe quando, na realidade, o documento pertencia a outra classe.

V_P representa a medida *Recall*, já anteriormente introduzida, que é usualmente referida como Sensibilidade neste contexto, enquanto $F_P$está relacionada com o conceito de Especificidade, que representa a taxa de verdadeiros negativos:

$$F_P = 1 - \text{Especificidade} \tag{2.32}$$

e,

$$\text{Especificidade} = \frac{VN}{FP + VN} \tag{2.33}$$

A curva ROC obtém-se fazendo variar o nível de decisão utilizado pelo classificador para atribuir, ou não, uma determinada classe a um documento. Para cada valor do nível de decisão obtém-se um par (V_P, F_P), de que são exemplos os pontos A, B, e C na Figura 2.8.6.1.

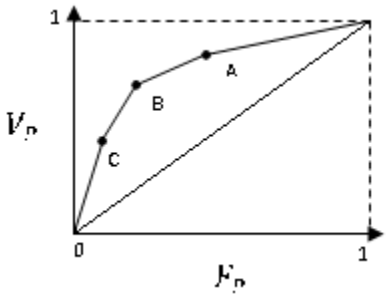

Figura 2.8.6. 1: Exemplo de uma Curva ROC

Quanto mais próxima do ponto (0,1) melhor o desempenho do classificador. A diagonal representa a situação do classificador aleatório.

A área abaixo da curva ROC é um indicador da exactidão do classificador. Quanto mais próxima de 1 melhor o desempenho do classificador, correspondendo o valor 1 ao classificador óptimo.

2.8.7. *Micro-Averaging*

Micro-Averaging é um método de cálculo da eficiência do classificador, muito utilizado na área de categorização de documentos (Cohen e Singer s.d.) (Debole e Sebastiani s.d.) (Thorsten n.d.) (How e Narayanan s.d.) (Zheng, Srihari e Srihari, A Feature Selection Framework for Text Filtering s.d.) (Zheng e Srihari, Optimally Combining Positive and Negative Features for Text Categorization s.d.) (Sebastiani e Debole, An analysis of the relative hardness of Reuters-21578 subsets s.d.) (Galavotti, Sebastiani e Simi s.d.). Considera-se *Micro-Averaging* baseada na Precisão (*Precision*), em *Recall* e em *F-Measure*. Esta medida auxiliará na análise comparativa dos resultados obtidos.

Considerando $C = \{c_1, c_2, \dots, c_N\}$ um conjunto de classes/categorias, e sendo VP_i, FP_i e FN_i, o número de Verdadeiros Positivos, o número de Falsos Positivos, e o número de Falsos Negativos, respectivamente, em relação à classe C_i, define-se a medida *Micro-Averaging* da seguinte forma:

Micro-Averaging Precision (*MicroP*):

$$MicroP = \frac{\sum_{i=1}^{N} VP_i}{\sum_{i=1}^{N} (VP_i + FP_i)} \tag{2.34}$$

Micro-Averaging Recall (*MicroR*):

$$MicroR = \frac{\sum_{i=1}^{N} VP_i}{\sum_{i=1}^{N} (VP_i + FN_i)} \tag{2.35}$$

Micro-Averaging F_1-Measure (*MicroF_1*):

$$MicroF_1 = \frac{2.MicroP.MicroR}{MicroP + MicroR}$$

(2.36)

Num problema de classificação com N classes, *MicroP* e *MicroR* podem ser determinados a partir da tabela *Micro-Averaging*, ou *Micro-Averaging Table* (MAT). Para construir esta tabela considera-se o problema de classificação binária para cada classe, ou seja constroem-se N tabelas, uma por classe, conforme exemplificado na tabela 2.8.7.1 para a classe c_i, em que VN_i representa o número Verdadeiros Negativos relativamente a essa classe.

	Classificador: Positivo	Classificador: Negativo
Real: Positivo	VP_i	FN_i
Real: Negativo	FP_i	VN_i

Tabela 2.8.7. 1: Classificação binária relativamente à classe c_i

A MAT obtém-se somando todas as tabelas relativas a todas as classes. A tabela 2.8.7.2 representa a MAT, em que *VP*, *FP*, e *FN*, representam o número total de Verdadeiros Positivos, o número total de Falsos Positivos, e o número total de Falsos Negativos, respectivamente.

	Classificador: Positivo	Classificador: Negativo
Real: Positivo	*VP*	*FN*
Real: Negativo	*FP*	*VN*

2.8.7. 1: Macro-Averaging Table (MAT)

As equações 2.34 e 2.35 podem ser rescritas dum modo mais compacto,

$$MicroP = \frac{VP}{VP + FP} \tag{2.37}$$

$$MicroR = \frac{VP}{VP + FN} \tag{2.38}$$

O cálculo destas medidas também pode ser efectuado a partir da Matriz de Confusão (MF), introduzida na secção 2.8.3, notando que

$$VP_i = MF(i, i) \tag{2.39}$$

$$FP_i = \sum_{\substack{k=1 \\ k \neq i}}^{N} MF(k, i) \tag{2.40}$$

e

$$FN_i = \sum_{\substack{k=1 \\ k \neq i}}^{N} MF(i, k) \tag{2.41}$$

e aplicando as equações 2.34 e 2.35.

Note-se que, no caso do presente estudo, a cada documento corresponde uma única classe, e portanto, aplicando a denotação das equações 2.37 e 2.38, tem-se

$$VP + FP = VP + FN = N\acute{u}mero\ total\ de\ documentos \tag{2.42}$$

o que implica

$$MicroP = MicroR = MicroF_1 \tag{2.43}$$

Tendo em conta a definição expressa na secção 2.8.4, e pelo que se referiu acima, no contexto deste estudo, a Exactidão também iguala as medidas expressas na expressão 2.43.

Capítulo 3

Resultados

3.1 Documentos de Treino e Teste

A colecção de documentos utilizados nesta dissertação de mestrado faz parte do *corpus Reuters*-21578 (Lewis s.d.).

O *corpus* da *Reuters* é muito conhecido e utilizado na área de categorização de documentos. Trata-se de uma colecção de artigos da *Reuters Newswire*[38] organizada manualmente por funcionários da *Reuters*, divulgada em 1987 e disponibilizada à comunidade científica em 1990. O conteúdo do conjunto de documentos que a compõem mereceu ser estudado mais atentamente. Consequentemente foram removidos documentos que tinham sido atribuídos a classes às quais estes não pertenciam, documentos que se encontravam repetidos, entre outras modificações. Este novo conjunto, conhecido por *Reuters*-22173, contém 22.173 documentos (Sebastiani e Debole, An analysis of the relative hardness of Reuters-21578 subsets s.d.).

Com *Reuters*-22173 esta colecção tornou-se numa referência padrão para comparação de trabalho entre investigadores (Cohen e Singer s.d.) . Em 1996 Finch e Lewis refinaram a colecção *Reuters*-22173 compondo um novo subconjunto da *Reuters*: *Reuters*-21578.

[38] *Reuters Newswire* é um jornal que contém notícias usualmente relacionadas com economia. É possível consultar via Web este jornal no sítio www.reuters.com.

No trabalho realizado nesta dissertação foram utilizadas quatro colecções, para tal foram utilizadas as colecções *Reuters*-21578 e *Reuters*-21578 *ModApté*.

A colecção *Reuters*-21578 *ApteMod* é uma colecção composta por um subconjunto dos documentos da colecção *Reuters*-21578. A colecção é composta por 10.788 documentos, sendo 7.769 documentos de treino e 3019 documentos de teste. Nesta versão da colecção *Reuters*-21578 *ModApté* foram retiradas ambiguidades e documentos que, devido ao seu conteúdo, originavam a criação de subconjuntos, da colecção *Reuters*-21578, que continham conjuntos de treino, e teste, diferentes. Passaremos a referir aos quatro conjuntos que utlizaram esta colecção por colecção R1, R2, R3 e R4.

A colecção R1 é composta por um subconjunto de documentos da colecção *Reuters*-21578. Os documentos considerados para esta colecção pertencem a um universo de 91 classes.

Com a colecção R1 foi possível analisar a capacidade de selecção de termos do 3M em relação a outras técnicas de selecção de termos, nomeadamente IG e CHI. E, com esta colecção estudou-se o comportamento destas técnicas, e dos classificadores, num universo de documentos pertencentes a 10, e a 91 classes.

Seguidamente referir-me-ei à colecção R1 de duas formas: R11 e R12. A colecção R11 contém o conjunto completo de documentos da colecção R1, sendo que nos documentos multi-classe foi considerado a classe dominante. A colecção R12 representa todos os documentos, da colecção R1, que pertencem a uma das 10 classes mais frequentes da colecção da *Reuters*-21578.

Com a colecção R1 foi possível realizar uma análise tendo em consideração três técnicas de selecção de termos, sete representações computacionais, e três classificadores. Apenas com esta colecção foi possível observar o comportamento de um classificador que não utiliza técnicas de selecção de termos (RIPPER) em relação a classificadores que utilizam técnicas de selecção de termos como SVM e K-NN.

No entanto com a colecção R1 não se poderá reter observações em relação a trabalho de outros autores (Debole e Sebastiani s.d.) (Thorsten n.d.) (Cohen e Singer s.d.) que nas suas experimentações tiveram em consideração todos os documentos da colecção da *Reuters*-21578. Devido a essa limitação, experimentou-se a técnica do terceiro momento com três colecções, R2, R3 e R4, assim, poder-se-á comparar os resultados obtidos com (Debole e Sebastiani s.d.) (Thorsten n.d.) (Cohen e Singer s.d.).

A colecção de documentos R1 foi subdividida em duas colecções:

- Conjunto R11: Os documentos que a compõem pertencem a um universo de 91 classes distintas.
- Conjunto R12: Os documentos que a compõem pertencem a um universo de 10 classes distintas. Sendo as 10 classes escolhidas aquelas que continham um maior número de documentos que as representassem.

Nas colecções R2, R3 e R4, tal como na colecção R12, foram tidas em consideração as 10 classes que continham um maior número de documentos para as representar.

Número de Documentos por Classe						
>100	>20 e <100	>0 e <20				= 0
acq	alum	barley	ipi	pet-chem	wpi	castor
crude	bop	carcass	iron-steel	platinum	yen	coconut-oil
earn	coffee	cocoa	jet	potato	zinc	copra-cake
interest	corn	coconut	jobs	propane		cotton-oil
money-fx	dlr	copper	l-cattle	rand		dfl
ship	gnp	cotton	lead	rapeseed		groundnut-oil
trade	gold	cpi	lei	reserves		lin-oil
	grain	cpu	livestock	retail		nz
	money-supply	dmk	lumber	rice		oat
	nat-gas	fuel	meal-feed	silver		palladium
	sugar	gas	naphtha	sorghum		palmkernel
	veg-oil	groundnut	nickel	soybean		rape-oil
	wheat	heat	nkr	soy-meal		rubber
		hog	nzdlr	soy-oil		rye
		housing	oilseed	strategi-metal		sun-meal
		income	orange	tea		sun-oil
		instal-debt	palm-oil	tin		sunseed

Tabela 3.1. 1: Informação do número de documentos por classe da colecção R11

Na tabela 3.1.1 é possível observar a média de documentos existentes por classe na colecção R11. A distribuição dos documentos pelas 91 classes não é proporcional, visto existirem classes que,

percentualmente, contêm muito mais documentos do que outras classes. O número de classes com o maior volume de documentos é inferior ao das classes que contêm uma quantidade inferior de documentos para as representar. Nos subconjuntos de documentos utilizados nesta dissertação existem classes que não contêm documentos que as representem, classes que contêm apenas entre 1 a 20 documentos, entre 20 a 100 documentos, e outras cujo número de documentos é superior a 100.

Com o subconjunto R1 foram realizados testes tendo em consideração documentos que pertenciam a um universo de 10 classes (R12), por um lado, e documentos que pertenciam a um universo de 91 classes (R11). A colecção utilizada para classificar o subconjunto R11, que tem em consideração 91 classes, contém 2876 documentos, e o subconjunto R12, que tem em consideração 10 classes, contém 2447 documentos. Na colecção R1 apenas foram seleccionados documentos que pertencem a uma única classe.

Na colecção R1 a ferramenta Weka seleccionou aleatoriamente o conjunto de documentos de treino e de teste. O conjunto de treino contém 80% dos documentos e os restantes 20% serão documentos de teste.

A colecção R1 apenas contém, aproximadamente, 27% dos documentos do *corpus* da *Reuters*-21578, sendo a selecção dos termos com maior relevância obtida com as técnicas IG (*Information Gain*) e CHI (*Chi-Square*) implementadas a nível da ferramenta Weka. Esta redução do *corpus* da *Reuters* impõe-se porque a ferramenta exige que se introduza a informação dos documentos da colecção num único ficheiro. Este ficheiro teria de conter a informação do conteúdo de todos os documentos, tendo em consideração todas as representações de documentos, ou seja, todos os termos. O tamanho do ficheiro para 10.788 documentos seria inexequível com a ferramenta Weka.

A colecção R1 possibilita observar os resultados obtidos com as diferentes técnicas de selecção de termos com diferentes classificadores.

No entanto, com a colecção R1 não se poderá comparar directamente os resultados obtidos com estudos em que toda a colecção da *Reuters*-21578 foi tida em consideração. Com vista a superar

esta limitação a técnica de selecção de termos 3M foi experimentada com diferentes colecções de documentos. As colecções R2, R3 e R4 apenas foram testadas com esta técnica porque o 3M, ao indicar os termos mais relevantes, não depende da ferramenta Weka. Permitindo assim determinar quais os termos mais relevantes diminuindo substancialmente a informação, por documento, introduzida no ficheiro utilizado pela ferramenta Weka.

A colecção R2 tem em consideração todos os documentos da colecção *Reuters*-21578. Os documentos que pertencem a mais de uma classe foram replicados.
Por exemplo, o documento *test*/14849 *interest money-fx* foi replicado nos documentos:

- *test*/14849 *interest*
- *test*/14849 *money-fx*

Nesta colecção o desempenho do classificador é consideravelmente prejudicado. Isto acontece porque mesmo quando um documento é correctamente classificado, se pertencer a mais de uma classe, os restantes documentos replicados serão incorrectamente classificados.

O ficheiro inserido na ferramenta Weka não permite que um documento esteja relacionado com mais de uma classe, ou seja, não possibilita a classificação de documentos multi-classe. Devido a esta limitação, na colecção R2, replicaram-se os documentos multi-classe. O objectivo seria de não excluir nenhum documento do conjunto inicial, tendo subjacente que o desempenho do classificador iria ser afectado.

A colecção R3 tem em consideração todos os documentos da colecção *Reuters*-21578. No caso dos documentos multi-classe estes foram renomeados para apenas ser tido em consideração a classe dominante do documento.
Por exemplo:

- `test/14849 interest money-fx → test/14849 interest`
- `test/14890 money-fx interest → test/14890 money-fx`

O desempenho do classificador também será afectado porque o conteúdo do documento também estará relacionado com outras classes que não estão a ser tidas em consideração. No entanto, por se considerar a classe dominante espera-se que não seja muito significativo.

Na colecção R4 foram excluídos da colecção *Reuters*-21578 os documentos multi-classe. Com este conjunto poder-se-á treinar, e testar, o classificador sem o "confundir" com documentos replicados por diferentes classes; ou documentos que continham informação relacionada com mais de uma classe.

Nas colecções R2, R3 e R4 os conjuntos de documentos de treino e teste são distintos. Foram considerados os documentos que pertencem às 10 classes com um número maior de documentos para as representar.

- Colecção R2 – 7.191 documentos de treino e 2.788 documentos de teste.
- Colecção R3 – 6.088 documentos de treino e 2.392 documentos de teste.
- Colecção R4 – 5.501 documentos de treino e 2.189 documentos de teste.

3.2. Resultados Experimentais

A tarefa de seleccionar os melhores termos, e treinar o classificador, para o conjunto de documentos com as 91 classes será mais exigente. Existem classes com poucos documentos para as representar, e por isso, considerou-se interessante observar a performance dos classificadores com as diferentes técnicas de selecção de termos. Com as colecções que apenas contêm documentos das 10 classes mais frequentes será possível observar se haverá uma melhoria de performance, e se a selecção das diferentes técnicas terá o mesmo impacto no classificador tanto para as 91 classes, como para as 10 classes. Outros autores também realizaram estudos tendo em consideração todas as classes, ou com subconjuntos das classes existentes na colecção *Reuters*-21578 (Debole e Sebastiani s.d.) (Sebastiani e Debole, An analysis of the relative hardness of Reuters-21578 subsets s.d.).

O número de documentos seleccionados do conjunto de documentos disponíveis na colecção da *Reuters*-21578 foi condicionado pela dimensão do ficheiro ARFF. Previamente à aplicação das técnicas de selecção de termos, sobre um conjunto de documentos, é gerado um documento em que são seleccionados todos os termos existentes no conjunto. E, para cada documento, é indicada a frequência de cada um desses termos. No caso da representação computacional por palavras, na colecção R11, são obtidas 14.984 palavras distintas. O formato do ficheiro ARFF exige que, para cada termo seja indicado, em cada documento, a sua frequência. Cada um destes documentos terá 14.984 entradas mais a informação da sua classe, contendo a colecção R11 2877 documentos. Por isso o ficheiro ARFF, com a representação computacional por palavras, além da indicação sequencial de cada termo, irá conter a seguinte informação:

- Número de documentos * (Número de termos+classe do documento):
  2877*(14.984+1)

Esta informação depende do número de documentos da colecção e, neste exemplo, adicionar um documento implica a entrada de 14.985 valores no ficheiro ARFF. Computacionalmente é possível ter em consideração mais documentos; no entanto este facto torna o ficheiro ARFF muito pesado para subsequente análise pela ferramenta Weka.

Na análise realizada, para as diferentes representações de documentos, existe uma variação do número de documentos utilizados na classificação. Tal acontece por existirem documentos em que não existe no seu conteúdo, pelo menos um termo, da representação computacional escolhida, no seu conteúdo. Existem documentos em que o seu conteúdo se descreve por:

Ficheiro da classe earn – 20979 [39]
19-OCT-1987 19-OCT-1987

Tabela 3.2. 1: Exemplo do conteúdo de um documento do corpus

São extraídas, do exemplo do documento descrito na tabela 3.2.1, três representações computacionais: 19; OCT; e 1987. Este documento não contém um único termo no caso das representações por multi-palavras ou sequências dos primeiros 4, 5 ou 6 caracteres. A numeração não foi considerada como uma candidata a termo. Por isso "OCT" será a única representação a ser considerada como um possível termo. Como não se trata de nenhuma multi-palavra, e por ser uma sequência inferior a 4 caracteres apenas será considerada um termo aquando da utilização da representação de documentos por palavras, em que o tamanho da sequência é irrelevante.

Documentos que não contenham pelo menos um termo, no seu conteúdo, não serão considerados para a classificação. E por isso existe uma variação do número de documentos de treino, e teste, utilizados por os classificadores com as diferentes representações computacionais.

[39] Cada ficheiro de teste e treino disponíveis na colecção *Reuters*-21578 (Lewis s.d.) estão associados a um número.

	Representação de Documentos	Número de Documentos	Número de Termos
91 Classes	Palavras	2877	14984
	Primeiros 4 Caracteres	2863	5564
	Primeiros 4,5 e 6 Caracteres	2863	18.805
	Multi-palavras	2863	14.000
	Primeiros 5 Carateres	2852	6585
	Pentagramas	2852	15.803
	Primeiros 6 Caracteres	2829	6656
10 Classes	Palavras	2447	13.222
	Primeiros 4 Caracteres	2434	5113
	Primeiros 4,5 e 6 Caracteres	2434	16.861
	Multi-palavras	2434	14.000
	Primeiros 5 Carateres	2423	5872
	Pentagramas	2423	14.122
	Primeiros 6 Caracteres	2400	5876

Tabela 3.2. 2: Número de documentos classificados por cada representação de documentos na colecção R11

Na tabela 3.2.2 é indicado o número de documentos que foram classificados por cada representação de documentos, e o número total de termos que foram identificados.

A representação por palavras é a única representação que engloba todos os documentos. Isto acontece porque não existem documentos sem nenhuma informação alfabética, e por isso, qualquer sequência de letras separadas por um caracter que não seja alfabético representará uma palavra. O mesmo não sucederá no caso das restantes representações.

As representações pelos primeiros 5-caracteres e pentagramas contêm o mesmo número de documentos classificados, uma vez que, independentemente do número de pentagramas que um documento possa conter, o primeiro pentagrama é sempre uma sequência de 5 caracteres. Para existir um documento que contenha uma sequência de 5 caracteres este será um pentagrama. Logo o número de documentos classificados será determinado pelos documentos que contenham pelo menos uma sequência de 5 caracteres, ou seja, um pentagrama. O mesmo acontece para os documentos classificados para os primeiros 4 caracteres, e o conjunto dos primeiros 4,5 e 6 caracteres. Para um documento conter uma sequência de 5 ou 6 caracteres terá necessariamente uma sequência de 4 caracteres; o contrário já não se aplica. Poderão existir documentos que contenham palavras com apenas 4 caracteres, ou menos. Por isso o número de documentos

classificados por ambas as representações será determinado pelos documentos que contenham palavras com pelo menos uma sequência de 4 caracteres.

Como foi referido no capítulo 2.1., o conjunto das multi-palavras foi extraída automaticamente pelo extractor de Aires et al. (Aires, Lopes e Silva s.d.). Para as técnicas de selecção de termos foram consideradas as primeiras 14.000 multi-palavras das 54.400 obtidas pelo extractor.

Relativamente à representação por multi-palavras observou-se que os documentos que não continham, pelo menos, uma multi-palavra correspondiam aos mesmos documentos que não continham, pelo menos, uma sequência com 4 caracteres. Nas 14.000 multi-palavras existem várias multi-palavras em que o número de caracteres de cada palavra, que a compõe, contém 4, ou menos, caracteres.

Os classificadores foram testados, para a colecção R1, com as diferentes técnicas de selecção de termos, e experimentados com vários conjuntos de termos. Observaram-se os resultados com os conjuntos de 50, 100, 300, 400, até 4500 termos que obtiveram maior pontuação com as diferentes técnicas de selecção de termos.

Na colecção R1 foram experimentadas as técnicas de selecção de termos CHI e IG, com dimensões variáveis do número de termos com maior relevância, com os classificadores K-NN e SVM. No caso do classificador SVM observou-se, aumentando o número de termos acima de 400, com excepção da representação por multi-palavras, não se observam melhorias significativas na performance do classificador. Se for utilizado um conjunto inferior o desempenho diminui, com o aumento do conjunto observa-se que o desempenho aumenta, mas com uma percentagem muito baixa. Quanto maior for o conjunto de termos seleccionado, mais tempo o classificador SVM necessita para gerar o classificador, e finalmente o classificar. Contudo, observou-se que seria necessário aumentar o conjunto de termos, acima dos 400, com vista a obter melhores resultados com a técnica do 3M. Com 2000 termos encontrou-se um equilíbrio entre a performance do classificador e as três técnicas de selecção de termos.

Na ferramenta Weka o classificador SVM foi parametrizado pelos parâmetros por defeito. Os documentos de treino foram normalizados e 1 para o nível de complexidade.

O classificador K-NN, para a colecção R1, retorna melhores resultados quando é utilizado um conjunto de termos que não seja muito elevado, aquando da sua utilização com diferentes representações de documentos e com as técnicas CHI, IG e 3M. Com K-NN quando se aumenta o conjunto de termos, acima de um limiar encontrado, a performance do classificador diminui consideravelmente. Por isso, optou-se por utilizar o conjunto de 400 termos para a experimentação com o classificador K-NN.

Como já foi referido no capítulo 2.5.7, o classificador K-NN, sendo k um valor inteiro, identifica qual a classe de um documento teste com base nas classes a que os k vizinhos mais próximos pertencem. A distância utilizada para obter a distância de um documento a outro foi a distância euclidiana. Na ferramenta Weka é possível parametrizar o cálculo da distância utilizada pelo classificador K-NN, e observou-se que para todas as classificações, independentemente da representação computacional, da técnica de selecção de termos escolhidas, do tamanho do conjunto de termos seleccionado, que o classificador K-NN retorna sempre melhores resultados quando a distância é pesada pelo seu inverso (Lavesson e Davidsson s.d.).

Actualmente ainda não existe um valor k que seja indicado em comum acordo entre os autores que realizaram estudos com o classificador K-NN. Autores como Huisman (Huisman s.d.), Shepperd e al. (Shepperd e Cartwright s.d.), e Myrtveit et al (Myrtveit, Stensrud e Olsson s.d.) sugerem valores de k entre 1 e 2. Shepperd e Cartwright (Shepperd e Cartwright s.d.) alertam para aquando da utilização de k=1 o algoritmo torna-se muito sensível a valores que poderão ter-se afastado do seu limiar. Ainda existem autores como (Batista e Monard s.d.) que indicam o valor de k=10 como o valor ideal para se obter um melhor desempenho com o classificador K-NN.

Para determinar o valor de k foi necessário testar e observar qual o valor que retornava a melhor performance. No caso da colecção R11, em que se teve em consideração as 91 classes, verificou-se que se obtinha melhores resultados quando se utilizava k=10. No entanto, no caso do da colecção R12, com as 10 melhores classes, quando se testa o classificador com valores de k superiores a 1 os resultados pioraram consideravelmente sendo, por isso utilizado k=1.

As experiências feitas com as colecções R2, R3 e R4, para o classificador K-NN, demonstraram que se teria de parametrizar o valor de k por 10. O desempenho do classificador diminui de dois a três pontos percentuais para valores de k inferiores a 10, para valores de k superiores a 10 o desempenho diminui a uma taxa mais elevada. O classificador retorna melhores resultados quando a distância aos documentos mais próximos é medida pela distância euclidiana, pesada pelo seu inverso, como já se tinha observado com a colecção R1.

As colecções R2, R3 e R4 foram apenas experimentadas com a técnica de selecção de termos 3M. Os classificadores K-NN, e SVM, foram experimentados com o conjunto de termos com maior relevância de dimensões de 400 a 6.000. Observou-se que a sua performance diminuía quando se optava por dimensões inferiores a 2000.

3.3. Pesos atribuídos pelas Técnicas de Selecção de Termos para a colecção R11

Neste subcapítulo serão descritos os termos com maior peso, para cada representação computacional, da colecção R11. O objectivo será observar o padrão existente entre os termos seleccionados por cada técnica de selecção de termos.

Nas tabelas que se seguem será possível visualizar quais os 14 termos com maior pontuação para cada representação de documentos: palavras; multi-palavras; primeiros 4, 5, e 6 caracteres, individualmente, e globalmente; e pentagramas, com cada uma das diferentes técnicas de selecção de termos. São indicados a sombreado os termos que são comuns, das 14 mais pontuadas, às técnicas do 3M, CHI e IG.

Terceiro	Momento
Termo	Peso
apr	7834,778
abbett	2289,1796
jun	2239,0745
a	2127,758
oct	1664,8745
ageements	1133,4362
abbott	380,76613
ago	346,414
alcoa	308,04675
algeria	244,38634
acquisittion	225,96834
ajusted	188,28275
and	166,57566
accepts	136,66574

Chi	Square
Termo	Peso
o	3297,6046
copper	3146,0633
cocoa	2707,0023
lt	2706,5645
coffee	2413,4265
sugar	2267,9882
vs	2255,9224
wheat	2195,4960
shr	2050,9754
net	2025,6939
cts	1938,0599
qtr	1848,0771
grain	1774,6233
corn	1751,6456

Information	Gain
Termo	Peso
lt	0,7201
vs	0,68
shr	0,6063
net	0,6027
cts	0,5698
qtr	0,5431
said	0,4359
the	0,3509
it	0,2928
to	0,2767
note	0,2534
revs	0,2522
inc	0,2463
a	0,2459

Tabela 3.3. 1: Pontuação obtida com Terceiro Momento, CHI e IG utilizando Palavras

Terceiro	Momento
Termo	Peso
aaron rents	10000
api says distillates	3143,5268
agreement on tariffs and trade	1978,2555
able to control production	1348,0919
about seven pct	660,8043
abbott laboratories	634,1662
ab astra	621,84854
adelaide river at mt	523,358
anz banking group says	417,9512
based on	374,88813
an official at	355,7636
allow a new issue	303,47888
comment on	272,20213
distillates off	235,3258

Chi	Square
Termo	Peso
roasted coffee	2528,6133
dlrs for coffee	2046,4907
cts vs	1903,4922
billion vs	1903,4922
p vs	1903,4922
mln vs	1903,4922
white sugar	1797,2619
unsold sugar	1797,2619
holly sugar	1797,2619
soft wheat	1707,5498
oper net	1568,8439
shr nil vs	1518,5887
mln credit vs	1518,5887
german feed wheat	1461,4161

Information	Gain
Termo	Peso
mln vs	0,5646
billion vs	0,5646
cts vs	0,5646
p vs	0,5646
oper net	0,4539
mln credit vs	0,4449
shr nil vs	0,4449
shr three cts vs	0,3884
shr seven cts vs	0,3884
shr four cts vs	0,3884
shr loss three cts vs	0,3671
shr loss eight cts vs	0,3671
oper shr	0,3508
for the	0,3317

Tabela 3.3. 2: Pontuação obtida com Terceiro Momento, CHI e IG utilizando Multi-palavras

Terceiro	Momento
Termo	Peso
aabe	10000,001
aaic	9398,694
abbe	8755,684
aaro	5738,1382
abbo	884,0974
agee	868,0099
aame	403,44844
acre	312,93774
augu	177,97459
boxe	168,09449
alge	167,43498
apri	149,76604
alco	148,72315
bill	142,4983

Chi	Square
Termo	Peso
copp	3130,7450
alum	2613,2787
coco	2560,4875
coff	2401,6596
suga	2256,9989
whea	2184,7755
tonn	1574,8288
said	1533,3955
crud	1466,3255
agri	1351,3962
gold	1226,6173
usda	1215,6850
barr	1210,9862
rate	1190,1001

Information	Gain
Termo	Peso
said	0,4386
revs	0,2541
acqu	0,2333
mths	0,2296
tonn	0,2126
rate	0,1971
note	0,1927
trad	0,1909
expo	0,1758
with	0,162
nine	0,1548
corp	0,1532
loss	0,1514
impo	0,1506

Tabela 3.3. 3: Pontuação obtida com Terceiro Momento, CHI e IG utilizando os primeiros 4 Caracteres

Terceiro	Momento	Chi	Square	Information	Gain
Termo	Peso	Termo	Peso	Termo	Peso
aabex	10000	coppe	3118,7093	acqui	0,2387
aaica	9920,6696	cocoa	2683,4711	tonne	0,212
abbet	6556,242	alumi	2603,2296	trade	0,2034
aaron	2950,3729	coffe	2392,4140	share	0,1968
ageem	2475,0216	sugar	2248,3001	expor	0,1819
archi	1797,7512	wheat	2176,3406	minis	0,1493
banpo	1112,0514	tonne	1584,8298	agree	0,1482
abbot	845,8162	crude	1460,4937	crude	0,1417
burnd	543,24017	agric	1399,5485	agric	0,1408
airse	478,1396	tarif	1114,9010	econo	0,1395
allia	413,2249	barre	1102,4894	price	0,1374
ander	359,30622	grain	1101,3098	produ	0,1327
edwar	275,64023	trade	1095,9788	rates	0,1309
alger	261,4795	expor	1070,9778	impor	0,1289

Tabela 3.3. 4: Pontuação obtida com Terceiro Momento, CHI e IG utilizando os primeiros 5 Caracteres

Terceiro	Momento	Chi	Square	Information	Gain
Termo	Peso	Termo	Peso	Termo	Peso
abbett	9961,549	copper	3093,5436	tonnes	0,1992
abando	4338,0833	alumin	2582,2233	export	0,1824
airsen	2199,9002	coffee	2373,0824	compan	0,1703
ageeme	1756,2298	tonnes	1604,0821	acquir	0,1642
anders	1479,1537	agricu	1445,8921	shares	0,1565
archit	1161,9294	barrel	1118,9218	minist	0,1495
abbott	967,1169	tariff	1105,6509	agricu	0,1477
banpon	696,83366	export	1061,0826	econom	0,1399
barric	624,10674	irania	893,2607	produc	0,1325
abatem	446,15455	attack	889,5323	import	0,1292
bulcan	369,78512	stabil	837,4114	prices	0,1193
burndy	339,99968	dollar	833,3925	offici	0,1184
canand	279,4443	import	830,8293	market	0,1164
chainw	267,31093	inflat	825,0591	barrel	0,1159

Tabela 3.3. 5: Pontuação obtida com Terceiro Momento, CHI e IG utilizando os primeiros 6 Caracteres

Terceiro Momento		Chi Square		Information Gain	
Termo	Peso	Termo	Peso	Termo	Peso
aabe	9999,999	copper	3130,7450	said	0,4386
aaic	9618,296	copp	3130,7450	revs	0,2541
abbe	5710,077	coppe	3130,7450	acqui	0,2384
aabex	5000,0046	cocoa	2693,8250	acqu	0,2333
aaica	4902,722	alum	2613,2787	mths	0,2296
aaro	4450,8812	alumin	2613,2787	tonn	0,2126
abbet	3634,6573	alumi	2613,2787	tonne	0,2115
aaron	3080,0097	coco	2560,4875	trade	0,2032
abbett	2665,7497	coffe	2401,6596	tonnes	0,1978
abbo	761,90853	coffee	2401,6596	rate	0,1971
agee	615,86566	coff	2401,6596	share	0,196
ageem	580,13763	suga	2256,9989	note	0,1927
ageeme	548,3271	sugar	2256,9989	trad	0,1909
abbot	435,0848	whea	2184,7755	export	0,1816

Tabela 3.3. 6: Pontuação obtida com Terceiro Momento, CHI e IG utilizando os primeiros 4, 5, e 6 Caracteres

Terceiro Momento		Chi Square		Information Gain	
Termo	Peso	Termo	Peso	Termo	Peso
aabex	9999,999	opper	3118,7093	acqui	0,2387
aaica	9892,0235	coppe	3118,7093	tonne	0,212
aaron	8398,6046	cocoa	2683,4711	trade	0,2034
abbet	3595,1496	alumi	2603,2296	onnes	0,1982
ageem	913,6228	lumin	2603,2296	share	0,1971
acled	913,4099	coffe	2392,4140	xport	0,1819
annoc	910,6461	offee	2392,4140	expor	0,1819
aacob	816,9488	sugar	2248,3001	ompan	0,1681
abbot	682,31	wheat	2176,3406	inist	0,1672
arlan	649,90716	onnes	1617,6925	cquir	0,1637
allas	474,88217	tonne	1584,8298	hares	0,1537
mason	435,5428	lture	1487,5690	minis	0,1493
bemis	403,221	crude	1460,4937	icult	0,147
airse	391,41347	icult	1458,0849	gricu	0,147

Tabela 3.3. 7: Pontuação obtida com Terceiro Momento, CHI e IG utilizando Pentagramas

Nas tabelas apresentadas é possível observar que, para todas as representações, a técnica CHI e IG têm muitos termos em comum. No conjunto dos 2000 melhores termos observou-se que os primeiros 14 termos, das técnicas IG e CHI, pertencem a ambos os conjuntos. Apesar de não terem sido pesados com a mesma relevância fazem parte do conjunto dos melhores 2000 termos.

A técnica do Terceiro Momento apresenta muitos termos que não existem no conjunto obtido com a técnica CHI e IG. Mas, nos conjuntos dos 400 e dos 2000 termos, o conjunto de termos seleccionado pela técnica 3M é muito distinto das restantes técnicas.

A pontuação atribuída a cada um dos termos, pela técnica do 3M, é realizada por um programa implementado na linguagem Java. Os termos são pontuados de acordo com o algoritmo do Terceiro Momento que foi descrito no capítulo 2.2.12. O programa Weka apenas irá auxiliar na classificação dos documentos com os diferentes classificadores, visto os termos já se encontrarem seleccionados e ordenados pela sua pontuação.

No caso das técnicas CHI e IG é o programa Weka que reordena e pontua os termos de acordo com a técnica seleccionada. Observou-se que em determinado momento a pontuação atribuída aos termos, antes de completar o conjunto dos 2000, foi nula. A ordenação dos termos, após esse instante, não é realizada de acordo com a técnica seleccionada pois esta sugere ser alfabética, apesar de, por vezes, ocorrer um termo que quebra essa ordem.

3.4. Resultados obtidos com a colecção R1

3.4.1. Resultados com o classificador SVM

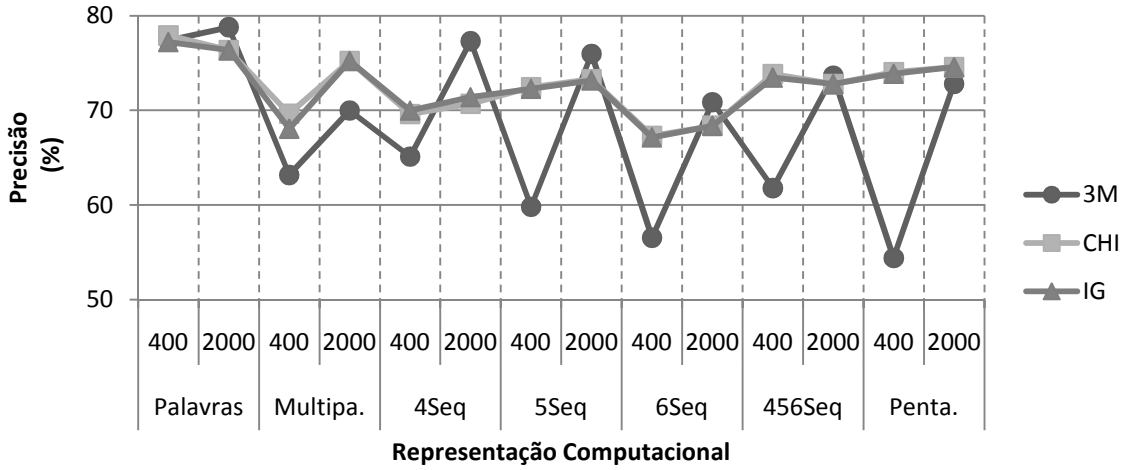

Figura 3.4.1. 1: Precisão obtida com o classificador SVM para cada uma das representações de documentos com as diferentes técnicas de selecção de termos com a colecção R11

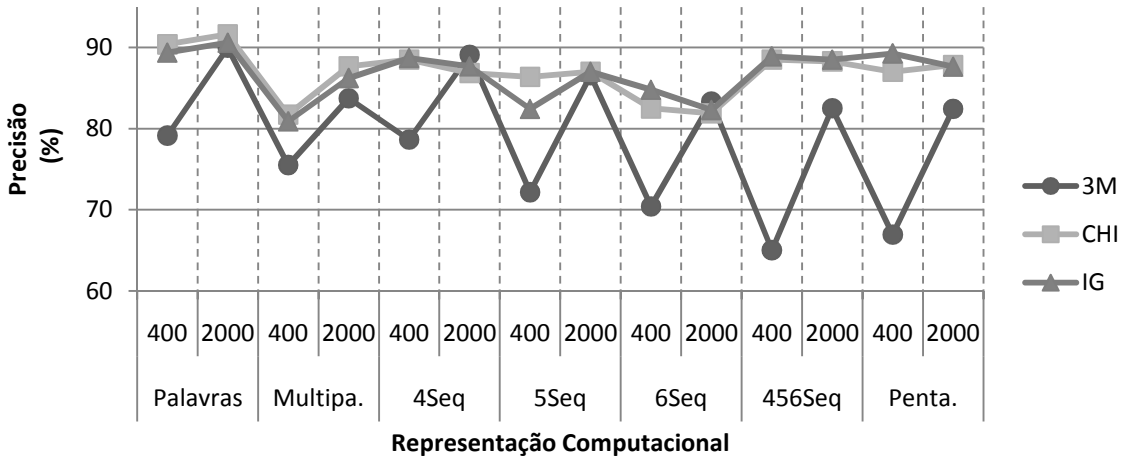

Figura 3.4.1. 2: Precisão obtida com o classificador SVM para cada uma das representações de documentos com as diferentes técnicas de selecção de termos com a colecção R12

Observando a figura 3.4.1.1. verifica-se que a precisão aumenta quando se utilizam os melhores 2000 termos ao invés de 400.

A técnica do terceiro momento é aquela em que se observam diferenças mais significativas aquando da variação do número de termos escolhido. Com algumas representações, como sequências dos primeiros 4 caracteres, ou primeiros 5 caracteres o classificador SVM, para a técnica de selecção de termos 3M, aumenta a sua precisão da mais baixa para a mais elevada.

Se for utilizado um conjunto de termos superior a 2000 observa-se que se mantém a mesma performance, ou que existe uma melhoria pouco significativa. Porém, se for utilizado um conjunto de 400 termos, ou um conjunto de tamanho inferior, a performance piora consideravelmente.

Observa-se que o classificador apresenta uma maior precisão, tanto para o subconjunto de 91 classes como para o de 10 classes, com a representação de documentos por palavras.

No subconjunto de 91 classes, a técnica do terceiro momento, quando é utilizado o conjunto dos 2000 termos com maior pontuação, demonstra para todas as representações, com excepção de multi-palavras, uma precisão superior ou equivalente à obtida com as outras técnicas de selecção de termos. No entanto, com o subconjunto de 10 classes, a precisão obtida com o terceiro momento é equivalente, ou inferior, à obtida com *chi-square* ou *information gain*. Estes resultados poderão indicar que a técnica do terceiro momento terá uma melhor capacidade de detectar termos mais relevantes de cada classe quando existem poucos documentos para as representar, e por isso, demonstra uma superioridade na classificação nessas situações. Quando existe uma maior percentagem de documentos por classe retorna uma precisão próxima da obtida com as restantes técnicas, como se pode observar na figura 3.4.1.2., com excepção da representação por o conjunto dos primeiros 4, 5 e 6 caracteres e pentagramas.

A precisão obtida com o classificador aumenta aproximadamente dez pontos percentuais, para todas as representações de documentos, quando se utiliza o subconjunto tendo em consideração apenas documentos com as 10 classes mais frequentes.

3.4.2. Resultados obtidos com o classificador KNN

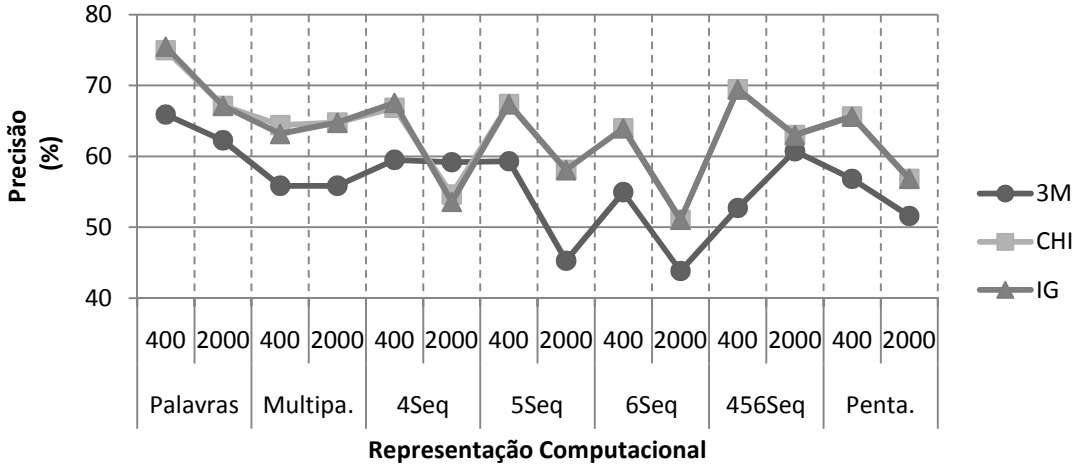

Figura 3.4.2. 1: Precisão obtida com o classificador KNN para cada uma das representações de documentos com as diferentes técnicas de selecção de termos com a colecção R11

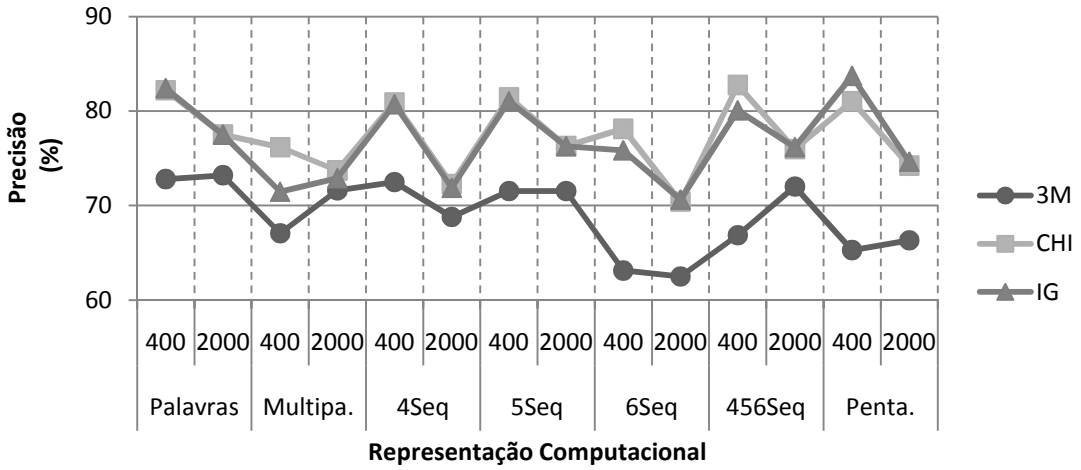

Figura 3.4.2. 2: Precisão obtida com o classificador KNN para cada uma das representações de documentos com as diferentes técnicas de selecção de termos para a colecção R12

Os resultados observados com o classificador K-NN, em relação aos resultados obtidos com o classificador SVM, apresentam uma diminuição de precisão de 2 a 21%. A maior diminuição de precisão apresentou-se nos resultados obtidos com a técnica do terceiro momento.

Com o classificar K-NN verifica-se, novamente, que os melhores resultados são obtidos para a representação de documentos por palavras. Contudo, com o subconjunto das 10 classes mais frequentes, as precisões mais elevadas são muito próximas entre várias representações, nomeadamente: por palavras; sequências dos primeiros 4 e 5 caracteres; conjunto dos primeiros 4,5 e 6 caracteres; e pentagramas.

É a técnica *information gain*, com a representação por pentagramas, no subconjunto das 10 melhores classes, que apresenta a melhor precisão. Sendo a segunda melhor precisão obtida com a representação pelo conjunto das cadeias dos primeiros 4, 5 e 6 caracteres, com uma diferença mínima em relação à primeira, obtida com a técnica *chi square*.

Com K-NN observa-se que os melhores resultados, com os subconjuntos de 91 classes e de 10 classes, são obtidos quando são tidos em consideração os 400 termos com maior pontuação. É possível observar nos gráficos 3.4.2.1. ou 3.4.2.2. que a precisão, com as diferentes representações, sendo a representação pelo conjunto de 4, 5 e 6 uma excepção, em ambos os casos, diminui, ou é equivalente, quando se aumenta o conjunto de termos.

Com a representação com o conjunto 4, 5 e 6 caracteres, tal como com as restantes representações, observa-se que com as técnicas *chi square* e *information gain* a precisão diminui com o aumento do conjunto de termos. Porém, com a técnica do terceiro momento a precisão aumenta com o aumento do conjunto de termos, sendo a precisão observada, com o conjunto de 400 e 2000 termos, inferior à que se observou com outras técnicas.

Observa-se na figura 3.4.2.1. e 3.4.2.2. que as precisões obtidas, com as técnicas *chi square* e *information gain*, se encontram muito próximas. Comportamento também observado com o classificador SVM.

3.5. Resultados obtidos com a colecção R2, R3 e R4

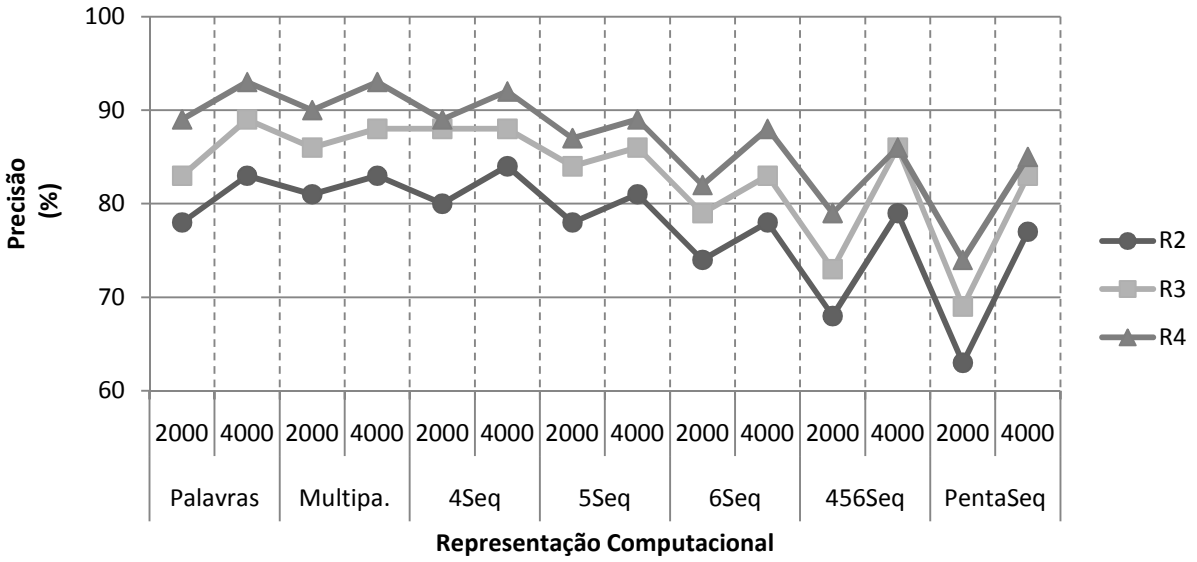

Figura 3.5. 1: Precisão obtida com o classificador SVM para as colecções R2, R3 e R4 com a técnica de selecção de termos 3M

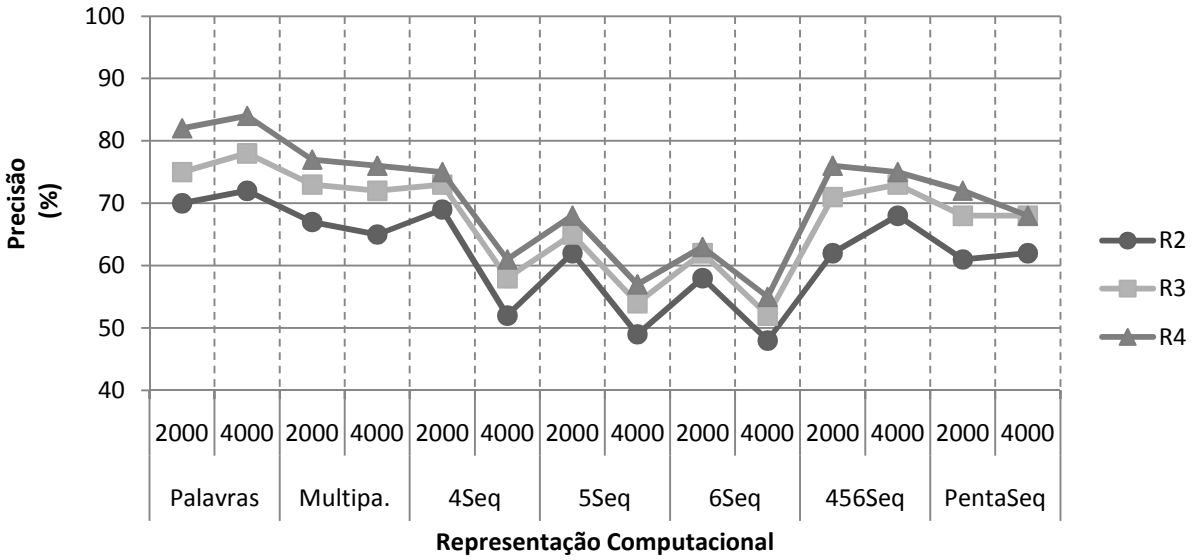

Figura 3.5. 2: Precisão obtida com o classificador K-NN para as colecções R2, R3 e R4 com a técnica de selecção de termos 3M

Com as colecções R2, R3 e R4, os classificadores K-NN e SVM têm um melhor desempenho quando se opta por seleccionar o conjunto dos termos com uma dimensão entre 2000 e 6000.

No caso da colecção R1, o classificador K-NN, diminuía o seu desempenho ao seleccionar um conjunto de termos com uma dimensão superior a 400. Nas restantes colecções o melhor desempenho foi obtido pelo conjunto de 2000 termos.

O classificador SVM, para a colecção R1, tendo em consideração um conjunto de termos com dimensão superior a 2000 observou-se que o desempenho era equivalente ao obtido ao obtido com a dimensão de 2000. Nas restantes colecções observou-se um aumento muito significativo no desempenho do classificador ao se optar pelo conjunto de termos com dimensão de 4000 invés de 2000.

Nas figuras 3.5.1. e 3.5.2. é possível observar que a precisão dos classificadores aumenta aquando da utilização da dimensão de 4000 temos. O melhor desempenho é obtido com a colecção R4.

O classificador SVM tem um melhor desempenho com as representações computacionais por palavras, multi-palavras e sequências de 4 caracteres. Seria espectável que o desempenho obtido com a representação computacional por 4, 5 e 6 caracteres fosse próxima da obtida com as sequências de 4 caracteres.

A dimensão do número de termos escolhido poderá ter sido a causa do desempenho obtido com a representação computacional por 4, 5 e 6 caracteres, com a colecção R4, ser aproximadamente seis pontos percentuais inferior à obtida com a representação computacional pelos primeiros quatro caracteres. Por isso, experimentou-se apenas, para a colecção R4, com os melhores 6000 termos.

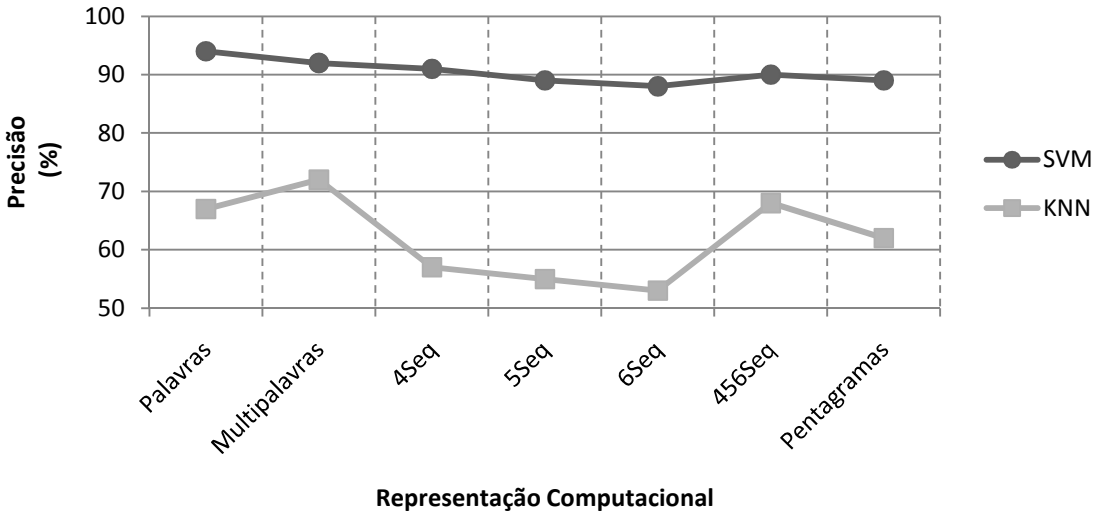

Figura 3.5. 3: Precisão obtida com o classificador SVM e KNN, com a colecção R4, para a técnica de selecção de termos do 3M e dimensão de 6.000 termos

Na figura 3.5.1 é possível observar os resultados obtidos pelos classificadores R2, R3 e R4, com o classificador SVM. Verifica-se que existe um aumento da precisão de 1 a 10 pontos percentuais quando se opta pelo conjunto dos melhores 4000 termos. Este aumento sugere que os 2000 termos para as colecções, com excepção da colecção R12, são insuficientes para representar as 10 classes.

As colecções R2, R3 e R4 contêm um conjunto de documentos de teste, e treino, muito superior à colecção R12. É possível que quando se considera o conjunto dos melhores 2000 termos com a colecção R12, nos documentos de teste, o número de documentos classificados incorrectamente seja muito inferior aos documentos classificados incorrectamente pelas restantes colecções.

Nas colecções R2, R3, e R4 as representações computacionais por palavra, multi-palavra e sequências de 4 caracteres são as que apresentam um melhor desempenho com o classificador SVM. Com estes resultados comprova-se a fiabilidade do desempenho obtido com o conjunto R12. Apesar de conter um número de documentos de treino e teste inferior o desempenho obtido com estas representações, com excepção da representação por multi-palavras, diferem entre 2 a 4

pontos percentuais. A representação por multi-palavras aumenta significativamente com a colecção R4 tendo uma exactidão superior à obtida com a colecção R12 de aproximadamente 8 pontos percentuais.

A representação computacional por sequências de 4 caracteres, com as colecções R2, R3 e R4, apresenta um melhor desempenho, de aproximadamente 3 e 5 pontos percentuais, em relação aos resultados obtidos com as representações computacionais por sequências de 5 e 6 caracteres. Seria espectável que os resultados obtidos com a representação computacional pelo conjunto de 4, 5 e 6 caracteres fosse próximo da representação pelas sequências de 4 cacteres, pois esta deveria conter os termos com maior peso em relação às restantes representações. O resultado obtido com o conjunto de 4,5 e 6 cacteres poderá ser devido ao facto de se utilizarem 4000 termos; e por isso, não se tendo em consideração termos essenciais que foram tidos em consideração pelo conjunto da representação computacional pelos primeiros 4 caracteres.

A colecção R4 retorna o melhor desempenho com os classificadores SVM e K-NN. Contudo, enquanto a precisão do classificador SVM aumenta com o aumento do conjunto de termos seleccionado pela técnica do 3M, o classificador K-NN diminui consideravelmente. Como já se tinha observado com a colecção R1 em que apenas se considerou 400 termos para o classificador K-NN, a determinada altura o classificador K-NN diminui a sua performance com o aumento da dimensão dos termos seleccionados.

Observou-se com a representação computacional por pentagramas, com o classificador K-NN, um aumento de 25 pontos percentuais com a colecção R4 quando se tem em consideração os melhores 4.000 termos. Não se encontrou uma justificação plausível para este resultado.

3.6. Resultados obtidos com o classificador RIPPER

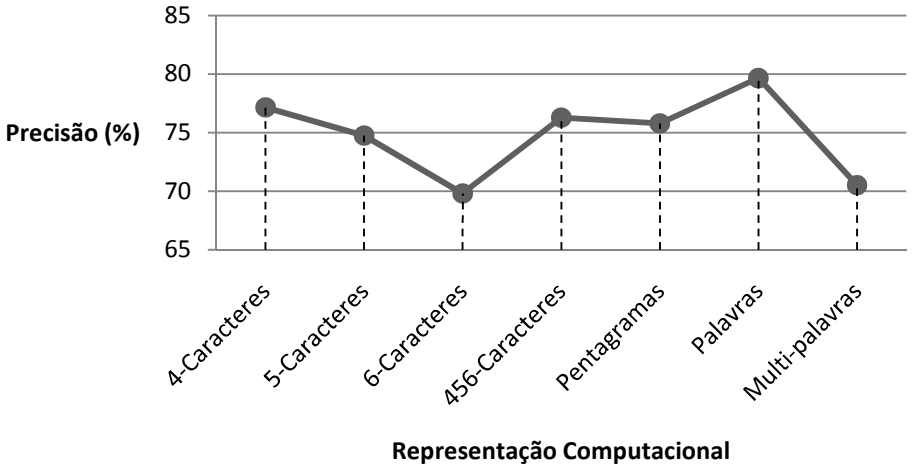

Figura 3.6. 1: Precisão obtida com o classificador RIPPER para cada uma das representações de documentos da colecção R11

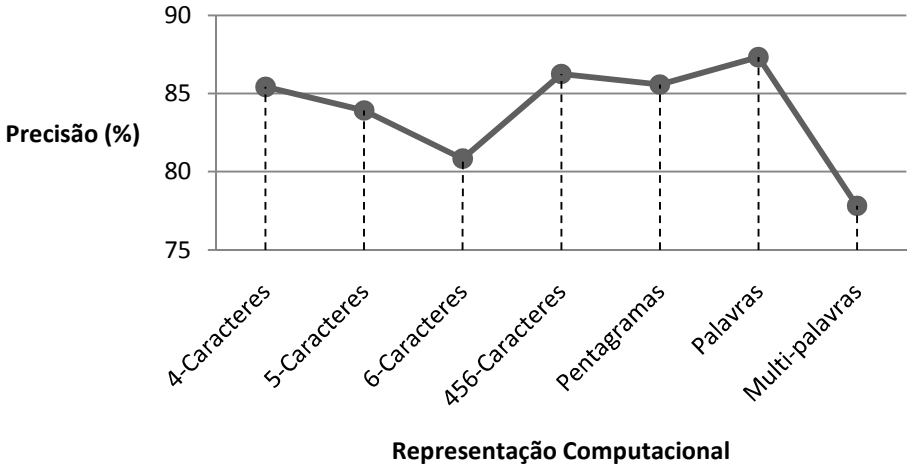

Figura 3.6. 2: Precisão obtida com o classificador RIPPER para cada uma das representações de documentos da colecção R12

Na figura 3.6.1. e 3.6.2. é possível observar a precisão obtida com o classificador RIPPER, com as colecções R11 e R12. A melhor precisão é obtida para a representação por palavras.

A representação por sequências dos primeiros 6 caracteres e multi-palavras são as representações, em ambos subconjuntos, que apresentam pior precisão.

O classificador RIPPER não foi testado com as colecções R2, R3 e R4 devido à dimensão do ficheiro ARFF. Como já foi indicado no capítulo 3.2. quanto maior for a dimensão do conjunto de termos maior será o ficheiro arff. Como o classificador RIPPER considera todos os termos para o conjunto de termos será o que tem a dimensão máxima para o conjunto de termos.

3.7. Resultados com SVM, K-NN e RIPPER

	RIPPER			SVM		K-NN	
	Exactidão (%)	Estatística Kappa	Técnica de Selecção de Termos	Exactidão (%)	Estatística Kappa	Exactidão (%)	Estatística Kappa
Palavras	79.65	0.73	3M	78.78	0.72	65.91	0.54
			CHI	76.35	0.69	74.96	0.66
			IG	76.35	0.69	75.48	0.67
Multi-Palavras	70.51	0.62	3M	69.98	0.62	55.85	0.40
			CHI	75.22	0.68	64.40	0.52
			IG	75.22	0.68	63.18	0.52
Sequências dos primeiros 4 caracteres	77.14	0.71	3M	77.31	0.71	59.51	0.45
			CHI	70.68	0.62	66.84	0.56
			IG	71.38	0.63	67.54	0.57
Sequências dos primeiros 5 caracteres	74.74	0.66	3M	75.96	0.68	59.30	0.42
			CHI	73.33	0.63	67.37	0.54
			IG	73.16	0.63	67.37	0.54
Sequências dos primeiros 6 caracteres	69.79	0.60	3M	70.85	0.61	54.95	0.38
			CHI	68.37	0.57	63.96	0.51
			IG	68.37	0.57	63.96	0.51
Sequências dos primeiros 4,5,6caracteres	76.27	0.70	3M	73.65	0.66	52.71	0.36
			CHI	72.77	0.64	69.46	0.60
			IG	72.77	0.64	69.46	0.60
Pentagramas	75.79	0.68	3M	72.81	0.63	52.63	0.37
			CHI	74.56	0.65	65.61	0.52
			IG	74.56	0.65	65.61	0.52

Tabela 3.7. 1: Exactidão e Estatística Kappa obtida por cada Classificador com o subconjunto da Reuters-21578 com colecção R11

Os resultados apresentados na tabela 3.7.1., para os classificadores SVM e K-NN, para as diferentes representações, são os que obtiveram melhores resultados. Para o classificador SVM os resultados apresentados são obtidos com o conjunto dos 2000 termos, e para K-NN com o conjunto dos 400 termos. O classificador RIPPER tem em consideração o conjunto completo. Por isso não foi necessário determinar qual seria o subconjunto de termos que se obteria melhores resultados.

	RIPPER			SVM		K-NN	
	Exactidão (%)	**Estatística Kappa**	**Técnica de Selecção de Termos**	**Exactidão (%)**	**Estatística Kappa**	**Exactidão (%)**	**Estatística Kappa**
Palavras	87.32	0.82	3M	89.98	0.86	72.80	0.61
			CHI	91.62	0.88	82.21	0.74
			IG	90.59	0.87	82.41	0.75
Multi-Palavras	77.82	0.69	3M	83.74	0.77	67.08	0.53
			CHI	87.68	0.83	76.18	0.66
			IG	86.24	0.81	71.46	0.59
Sequências dos primeiros 4 caracteres	85.42	0.80	3M	89.12	0.84	72.48	0.59
			CHI	86.86	0.81	80.90	0.72
			IG	87.68	0.82	80.70	0.72
Sequências dos primeiros 5 caracteres	83.92	0.76	3M	86.60	0.79	71.57	0.56
			CHI	87.01	0.80	81.44	0.71
			IG	87.01	0.80	81.03	0.70
Sequências dos primeiros 6 caracteres	80.83	0.72	3M	83.33	0.76	63.13	0.46
			CHI	81.86	0.73	78.13	0.67
			IG	82.29	0.74	75.83	0.64
Sequências dos primeitos 4,5,6caracteres	86.24	0.81	3M	82.51	0.75	66.87	0.52
			CHI	88.30	0.83	82.75	0.76
			IG	88.50	0.84	80.08	0.71
Pentagramas	85.57	0.78	3M	82.44	0.75	65.29	0.49
			CHI	87.84	0.82	81.03	0.71
			IG	87.63	0.81	83.71	0.75

Tabela 3.7. 2: Exactidão e Estatística Kappa obtida por cada Classificador com o subconjunto da Reuters-21578 com colecção R12

A tabela 3.7.2. tal como a tabela 3.7.1. tem em consideração o conjunto dos 2000 termos com maior pontuação para o classificador SVM, e o conjunto dos 400 termos com maior pontuação para o classificador K-NN.

Os resultados apresentados, na tabela 3.7.1., para o classificador K-NN foram obtidos com 10 para o valor de K. Na tabela 3.7.2., que corresponde ao subconjunto da Reuters-21578 com as 10 classes mais frequentes, o valor de K utilizado foi 1.

		Conjunto de 2000 Termos		Conjunto de 4000 Termos		Conjunto de 6000 Termos	
	Colecção	**Exactidão (%)**	**Estatística Kappa**	**Exactidão (%)**	**Estatística Kappa**	**Exactidão (%)**	**Estatística Kappa**
Palavras	R2	78.01	0.71	83.11	0.78		
	R3	83.42	0.75	88.51	0.83		
	R4	88.61	0.80	92.78	0.88	93.80	0.89
Multi-Palavras	R2	81.41	0.75	82.22	0.76		
	R3	86.34	0.80	88.19	0.83		
	R4	89.58	0.84	91,37	0.86	91.69	0.87
Sequências dos primeiros 4 caracteres	R2	80.40	0.74	83.83	0.79		
	R3	87.73	0.82	87.93	0.82		
	R4	89.42	0.83	91,64	0.87	91.23	0.86
Sequências dos primeiros 5 caracteres	R2	78.18	0.71	80.62	0.74		
	R3	83.83	0.76	85.86	0.79		
	R4	86.63	0.71	88.96	0.83	88.66	0.82
Sequências dos primeiros 6 caracteres	R2	73.50	0.65	78.39	0.71		
	R3	79.01	0.69	83.21	0.76		
	R4	81.63	0.71	87.90	0.81	87.76	0.84
Sequências dos primeitos 4,5,6caracteres	R2	67.51	0.55	78.91	0.72		
	R3	72.74	0.58	85.99	0.79		
	R4	78.69	0.65	85.98	0.78	90.17	0.86
Pentagramas	R2	62.98	0.50	77.27	0.70		
	R3	69.25	0.54	82.90	0.75		
	R4	73.79	0.58	84.90	0.76	88.59	0.82

Tabela 3.7. 3: Exactidão e Estatística Kappa obtida pelo classificador SVM e a técnica de selecção de termos 3M com as colecções R2, R3 e R4

Na tabela 3.7.3 observa-se que o desempenho do classificador SVM aumenta com a dimensão do conjunto de termos. As representações computacionais por palavra, multi-palavra, sequências de 4 caracteres, e sequências de 4, 5 e 6 caracteres são as que retornam a exactidão mais elevada.

A dimensão de 6.000, para o conjunto dos termos com maior pontuação, comprovam a necessidade de um número mais elevado de termos com vista a aumentar a exactidão obtida com a representação por sequências de 4, 5 e 6 caracteres. Trata-se da representação que teve o aumento da exactidão mais significativa, comprovando a necessidade de se considerar um maior número de termos para que esta representação computacional obtenha um desempenho próximo do obtido com as sequências de 4 caracteres.

		Conjunto de 2000 Termos		Conjunto de 4000 Termos		Conjunto de 6000 Termos	
	Colecção	**Exactidão (%)**	**Estatística Kappa**	**Exactidão (%)**	**Estatística Kappa**	**Exactidão (%)**	**Estatística Kappa**
Palavras	R2	70.01	0.59	71.60	0.63		
	R3	75.16	0.62	77.90	0.67		
	R4	81.74	0.67	84.44	0.72	66.85	0.30
Multi-Palavras	R2	67.48	0.56	64.83	0.51		
	R3	73.43	0.60	71.81	0.57		
	R4	77.64	0.65	76.41	0.61	72.01	0.52
Sequências dos primeiros 4 caracteres	R2	69.01	0.57	52.38	0.29		
	R3	73.36	0.59	57.70	0.29		
	R4	74.93	0.57	60.73	0.28	56.80	0.19
Sequências dos primeiros 5 caracteres	R2	61.94	0.46	49.12	0.24		
	R3	65.42	0.45	53.58	0.21		
	R4	67.58	0.43	56.66	0.20	54.69	0.15
Sequências dos primeiros 6 caracteres	R2	57.92	0.41	48.01	0.23		
	R3	61.65	0.40	51.76	0.19		
	R4	62.59	0.36	54.62	0.17	52.68	0.13
Sequências dos primeitos 4,5,6caracteres	R2	62.32	0.48	67.92	0.56		
	R3	70.54	0.56	72.78	0.58		
	R4	75.81	0.61	74.58	0.57	68.15	0.44
Pentagramas	R2	62.98	0.50	61.92	0.46		
	R3	68.42	0.53	66.71	0.47		
	R4	72.00	0.55	67.90	0.44	61.79	0.31

Tabela 3.7. 4: Exactidão e Estatística Kappa obtida pelo classificador K-NN e a técnica de selecção de termos 3M com as colecções R2, R3 e R4

Na tabela 3.7.4 observa-se que o classificador K-NN, ao contrário do classificador SVM, diminui o seu desempenho com o aumento da dimensão dos termos com maior pontuação.

O melhor desempenho foi obtido, com o conjunto dos melhores 2000 termos, pelas representações computacionais por palavras; multi-palavras; sequências de 4 caracteres; e sequências de 4, 5 e 6 caracteres. Sendo as mesmas representações computacionais que obtiveram melhor desempenho com o classificador SVM.

Observa-se que tal como na colecção R1, o classificador K-NN para as colecções R2, R3 e R4 apresenta um melhor desempenho com um conjunto de termos inferior, em constrate com o classificador SVM.

3.8. Desempenho por Classe

Nas secções anteriores toda a informação sobre o desempenho do classificador é obtida através dos resultados alcançados para todas as classes. No caso da precisão foi realizada uma média ponderada tendo em consideração a precisão obtida por cada classe e o número de documentos dessa classe existentes no conjunto teste.

O desempenho global, em que são tidas em consideração as 91 (e 10) classes do *corpus* da *Reuters*-21578, obtida pelo classificador, poderá ser afectada pela existência de classes em que o classificador teve um fraco desempenho em relação a outras classes.

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.784	0.968	0.866	0.956	livestock	0.667	0.667	0.667	0.991
alum	1	1	1	1	money-fx	0.55	0.458	0.5	0.904
bop	0.4	0.667	0.5	0.994	money-sup	0.4	0.4	0.4	0.941
carcass	0	0	0	0.902	nat-gas	0	0	0	0.989
cocoa	0	0	0	0.897	nzdlr	0	0	0	0.995
coffee	0.75	1	0.857	0.999	orange	1	0.333	0.5	0.997
copper	1	0.4	0.571	0.968	pet-chem	0	0	0	0.968
corn	0	0	0	0.983	platinum	0	0	0	0.411
cotton	1	1	1	1	potato	0	0	0	0.104
cpi	0.75	0.6	0.667	0.992	reserves	0.333	0.25	0.286	0.896
crude	0.667	0.609	0.636	0.909	rice	0	0	0	0.988
dlr	1	0.143	0.25	0.984	ship	0.615	0.571	0.593	0.881
earn	0.966	0.958	0.962	0.978	soybean	0	0	0	0.734
fuel	1	0.333	0.5	0.681	soy-meal	0	0	0	0.995
gas	0.5	1	0.667	0.999	soy-oil	0	0	0	0.801
gnp	0.5	0.667	0.571	0.994	strat-metal	0	0	0	0.611
gold	0.333	1	0.5	0.998	sugar	1	0.5	0.667	1
grain	0.333	0.333	0.333	0.732	tin	0	0	0	0.935
heat	0	0	0	1	trade	0.6	0.706	0.649	0.97
interest	0.515	0.68	0.586	0.966	veg-oil	0.5	0.333	0.4	0.937
ipi	1	0.2	0.333	0.982	wheat	0.75	0.75	0.75	0.988
iron-steel	0.667	0.5	0.571	0.97	wpi	1	0.5	0.667	1
jobs	0.5	0.5	0.5	0.983	yen	0	0	0	0.998
l-cattle	0	0	0	0.989	zinc	1	1	1	1
lei	0.5	0.5	0.5	0.997					

Tabela 3.8. 1: Performance por classe com representação de documentos por Palavras (SVM e 3M)

A tabela acima apresentada corresponde à tabela A.1.1, que se encontra no apêndice A, e corresponde à classificação realizada com o classificador SVM, com a técnica do terceiro momento, para representação dos documentos por palavras, com a colecção R11. No apêndice A, B e C, é possível consultar a performance por classe de cada um dos classificadores, por cada representação computacional, e, no caso dos classificadores SVM e K-NN, com cada técnica de selecção de termos.

Na tabela 3.8.1 foram retiradas as classes que não foram atribuídas a nenhum documento, erradamente ou correctamente. Esta situação ocorre quando existem classes que apesar de fazerem parte do corpus não tinham nenhum documento para as representar. Por isso não ocorre a situação de surgirem documentos classificados erradamente como se pertencessem a essa classe. Foram indicadas a sombreado as classes em que existiam documentos no *corpus* de teste que lhes pertenciam, mas em que nenhum foi correctamente classificado. Quando não existe, pelo menos, um documento de uma classe correctamente classificado a medida precisão e recall são nulas, logo, consequentemente também será a medida f-measure.

Observando a tabela 3.8.1., verifica-se que foi com a classe earn que se obteve melhor performance. Para o classificador SVM, com todas as técnicas de selecção de termos, e representações computacionais, a classe *earn* foi também a que obteve uma melhor performance. É de ter em consideração que se trata da classe dominante no número de documentos que existem para a representar.

A classe gold é um exemplo de uma situação em que o recall é muito elevado e a precisão é baixa. Esta situação ocorre quando existem documentos pertencentes a outra classe e que foram incorrectamente classificados como pertencente à classe gold, falsos positivos, e por isso, a precisão diminui. Mas poderá ocorrer o caso de não existirem falsos negativos, ou seja, documentos que pertencentes à classe gold e tenham sido classificados como pertencentes a outra classe, e por isso o recall será elevado. Na medida f-measure será possível observar que estas duas medidas são tidas em consideração, e por isso, a classe gold tem 0.5 com a medida f-measure apesar de apresentar 1 com recall.

Na tabela 3.7.1. a exactidão e kappa obtidos com o classificador SVM, para a representação por palavras, e utilizando a técnica do terceiro momento, são respectivamente: 78.7826 e 0.7226. Na tabela 3.8.1. observa-se com as classes earn e acq, que o desempenho encontra-se entre 86% e 96%. Pode-se concluir que a performance global do classificador se encontra muito afectada pela baixa performance de algumas classes.

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.936	0.954	0.945	0.968	146	1	2	0	2	1	0	1	0	0
earn	0.970	0.970	0.970	0.979	3	194	1	0	1	1	0	0	0	0
money-fx	0.733	0.815	0.772	0.944	1	1	22	1	0	0	2	0	0	0
grain	0.667	0.857	0.750	0.995	0	0	0	6	0	0	0	0	1	0
crude	0.800	0.667	0.7257	0.953	3	2	0	0	24	0	1	4	2	0
trade	0.842	0.842	0.842	0.956	1	2	0	0	0	16	0	0	0	0
interest	0.833	0.714	0.769	0.984	0	0	5	0	0	0	15	0	0	0
ship	0.583	0.636	0.609	0.980	1	0	0	0	3	0	0	7	0	0
wheat	0.583	0.700	0.636	0.982	1	0	0	2	0	0	0	0	7	0
corn	1	0.600	0.75	1	0	0	0	0	0	0	0	0	2	3

Tabela 3.8. 2: Performance obtida por classe e Matriz de confusão para Palavras (SVM e 3M)

A tabela acima apresentada corresponde à tabela D.1.1, que se encontra no apêndice D, e descreve a classificação realizada com o classificador SVM, com a técnica do terceiro momento, para representação dos documentos por palavras, com a colecção R12 em que se tem em consideração as 10 classes mais frequentes. No apêndice D, E e F é possível consultar a performance por classe de cada um dos classificadores, por cada representação computacional, e, no caso dos classificadores SVM e K-NN, com cada técnica de selecção de termos. No apêndice G e H é possível consultar a performance por classe, com o classificador SVM e K-NN, por cada representação computacional, com a técnica do terceiro momento, para a colecção R4.

Na tabela 3.8.2 é possível observar a performance por classe, e a matriz de confusão. Na matriz de confusão observa-se que as classes que contêm menos documentos de teste são as classes que têm a performance mais baixa. O que poderá indicar que a precisão obtida com o classificador, e com a técnica do terceiro momento, será entre 94% e 97% se existirem mais documentos representativos de cada classe. Com a colecção R4, em que foi tido em consideração um número

mais elevado de documentos, o desempenho do classificador foi de aproximadamente 94%. Os resultados com a colecção R4 comprovam a necessidade de existir um número mais elevados de documentos por classe.

3.9. Análise dos resultados obtidos em relação a outros autores

Neste capítulo comparam-se os resultados obtidos, com estudos apresentados por outros autores. O objectivo é comprovar, em relação a trabalhos de referência, que a técnica do terceiro momento tem a capacidade de superar, ou igualar, estas técnicas já muito conhecidas. Relembra-se que no capítulo 2.8.7 foi referido que no contexto deste estudo, a medida *Micro-Average* é equivalente à exactidão apresentada nos resultados obtidos com as colecções R1, R2, R3 e R4.

Muitos dos autores que se irão referir aplicaram sobre a colecção da *Reuters*-21578 técnicas que reduzem dimensionalidade da colecção ao utilizar listagens de palavras não interessantes (*stop-words*), ou técnicas como radicalização ou lematização. Como foi referido no capítulo 2.2 estas técnicas são dependentes da língua. No trabalho realizado nesta dissertação não foi aplicada nenhuma técnica que acarretasse dependências línguisticas.

No trabalho de Joachims (Thorsten n.d.) são experimentados vários classificadores, entre estes encontra-se os classificadores SVM e K-NN, com a colecção *Reuters*-21578. Foi aplicado um pré-processamento de dados com as técnicas de radicalização e uma lista de palavras não interessantes. *Information Gain* foi a técnica de selecção de termos escolhida foram seleccionados os melhores 1000 termos. Os termos correspondem à representação computacional por palavra.

Na experimentação realizada com o classificador K-NN a técnica de selecção de termos 3M, tendo em consideração os melhores 4.000 termos da colecção R4, atingiu 84.4% de exactidão para a representação por palavras. O valor obtido valor obtido por Joachims (Thorsten n.d.) foi de 82.3%.

Os resultados obtidos com o classificador SVM, para a colecção R4 e os melhores 4.000 termos, para todas as representações computacionais, com excepção da representação por sequências 456 caracteres e pentagramas, supera os cerca de 86% obtido no trabalho de Joachims (Thorsten n.d.). A representação computacional das sequências de 456 e pentagramas apenas apresentam 0,4% e

1,5%, respectivamente, de desempenho inferior ao obtido por Joachims. O desempenho obtido pela representação por palavras tem um desempenho superior na ordem de 6,4%.

Debole e Sebastiani (Debole e Sebastiani s.d.) realizaram experimentações com o classificador SVM para três técnicas de selecção de termos: *Chi-Square*, IG (*Information Gain*) e GR (*Gain Ratio*). Os autores tiveram em consideração, para cada técnica de selecção, os termos de forma localizada e global. Como foi indicado no capítulo 2.2. os termos podem ser considerados localmente em que cada classe é representada por um conjunto único de termos, ou globalmente onde os termos podem ser partilhados entre as classes.

No trabalho realizado em (Debole e Sebastiani s.d.) foram retiradas da colecção da *Reuters*-21578 palavras com base numa listagem de palavras não interessantes e utilizada uma técnica de radicalização.

Debola e Sebastiani indicam que obtêm melhores resultados com as técnicas GR e *Chi-Square* ambas consideradas globalmente. Quando se tem em consideração as 10 classes mais frequentes, o melhor desempenho é de aproximadamente 92.4%, para as 90 classes será de aproximadamente 87%.

Os resultados obtidos com a colecção R11 e a representação computacional por palavras são cerca de 9% inferiores à obtida pelos autores (Debole e Sebastiani s.d.). Estes resultados poderão ter sido influenciados pelo facto de se estar a considerar um número de classes muito elevado para o número baixo de documentos existentes, por classe, para as representar. A colecção R12, que apenas se tem em consideração as 10 classes com um maior número de documentos, já se aproxima dos resultados obtidos pelos autores, sendo apenas cerca de 2% inferior.

Com o classificador SVM, para a colecção R4, utilizando os melhores 4000 e 6000 termos, retorna um desempenho equivalente ao obtido por (Debole e Sebastiani s.d.), variando apenas na ordem de 1% acima ou abaixo. Estes resultados são obtidos pelas representações computacionais por palavras, multi-palavras, sequências de 4 caracteres e sequências de 456 caracteres.

Cohen e Singer (Cohen e Singer s.d.) experimentaram o classificador RIPPER com várias colecções diferentes, entre estas encontra-se *Reuters*-21578. Os autores utilizaram a representação computacional por palavras e foi utilizada uma lista de palavras não interessantes.

Cohen e Singer obtiveram um desempenho de aproximadamente 81.9% com a colecção da *Reuters*-21578 tendo em consideração 93 classes. O desempenho obtido no estudo realizado para a colecção R11, em que se tem em consideração 91 classes, com a representação computacional por palavras foi de 79.65%.

Como se verificou no parágrafo, anterior os resultados obtidos com a colecção R11, com o classificador RIPPER, foi aproxidamente 2% inferior aos resultados obtidos por Cohen e Singer. Na colecção R11, apesar de se ter em consideração 91 classes, existiam muitas classes com um número muito baixo, ou até mesmo inexistente, de documentos para as representar nos documentos de treino, e/ou nos documentos de teste. Seria espectável ao se ter em consideração a colecção completa da *Reuters*-21578 o resultado obtido seria equiparável ao obtido no trabalho de Cohen e Singer.

No estudo realizado com a colecção R11 observou-se que a representação computacional por sequências de 4 caracteres; sequências de 5 caracteres; sequências de 4, 5 e 6 caracteres; e finalmente, pentagramas obtiveram um desempenho muito próximo do obtido com a representação computacional por palavras.

A representação computacional por sequências por sequências de 4 caracteres visava prescendir de técnicas como lematização e radicalização. Com o classificador SVM observou-se que se obteve um bom desempenho, no entanto o classificador RIPPER esteve abaixo do que se pretendia. Ter-se-á de considerar que o classificador RIPPER apenas foi testado com a colecção R11, e R12, que continham uma proporção de documentos muito inferior aos documentos pertencentes à colecção *Reuters*-21578, e *Reuters*-21578 *ModApté*.

No trabalho de Gonçalves e Quaresma (Gonçalves e Quaresma s.d.), para a colecção da *Reuters* tendo em consideração as 10 classes mais frequentes, obtiveram um desempenho de 95%. Neste estudo (Gonçalves e Quaresma s.d.) o classificador SVM foi experimentado com três técnicas de

selecção de termos (*term frequecy*, *mutual information* e *gain ratio*), e posteriormente pesadas por quatro medidas de *term weighting*.

No trabalho realizado nesta dissertação os termos foram automaticamente normalizados pelo classificador SVM da ferramenta Weka, no entanto não foi previamente aplicada nenhuma medida de *term weighting*. Contudo, pode-se observar na colecção R4, em que se tem em consideração o conjunto de 6000 termos, para a representação computacional por palavras o desempenho obtido pelo classificador SVM é apenas, aproximadamente, 2% inferior ao obtido por Gonçalves e Quaresma, comprovando novamente, a conformidade dos resultados obtidos com a técnica do terceiro momento.

Capítulo 4

Conclusão

Nesta dissertação é apresentado um estudo na área de categorização automática de documentos. Actualmente existe uma necessidade de continuamente melhorar as técnicas de categorização porque se está numa era em que os avanços tecnológicos possibilitaram um aumento da velocidade disponibilização virtual de documentação, requerendo, por isso, para contrabalançar essa maior disponbilização de documentação, técnicas, de preferência independentes da lingua, com capacidade de rapidamente analisar e categorizar o conteúdo desta informação.

Apresentou-se uma implementação da técnica de selecção de termos Terceiro Momento na área de categorização de documentos. Esta técnica foi recentemente proposta pelo Professor Joaquim F. da Silva em conversa pessoal. O estudo realizado teve por objectivo testar esta técnica na área de categorização automática de documentos, tendo simultaneamente em consideração outras técnicas de selecção de termos, diferentes representações computacionais e, finalmente, mais de um classificador.

Na análise da capacidade de selecção de termos do 3M os testes foram realizados sobre uma colecção frequentemente utilizada na área de categorização: *Reuters*-21578 *ModApté* (Cohen e Singer s.d.) (Debole e Sebastiani s.d.) (Thorsten n.d.).

Na colecção da *Reuters*-21578 as classes não se encontram uniformememnte distribuídas, e observa-se que o desempenho dos algoritmos de selecção de termos, e de classificação, são sensíveis à distribuição de documentos pelas classes. O desempenho do classificador é consideravelmente superior com as colecções em que apenas se consideram os documentos que pertencem às 10 classes mais frequentes, relativamente ao desempenho equivalente obtido para o conjunto em que se têm em consideração 91 classes.

Na presente tese foi realizado um estudo comparativo de três classificadores: RIPPER, K-*Nearest Neighbour* e *Support Vector Machines*, aplicados à classificação de textos de acordo com o seu conteúdo temático. Com o objectivo de comparar o desempenho destes classificadores, no trabalho realizado, foram utilizadas sete representações computacionais dos textos utilizados no treino dos classificadores, e na fase de teste dos mesmos, nomeadamente, a representação por palavra, por multi-palavra, pelos primeiros 4, 5, e 6 caracteres das palavras dos textos, considerados individualmente e globalmente, e, finalmente, pelos pentagramas de caracteres.

Pretendeu-se também estudar o desempenho dos classificadores referidos, tendo em linha de conta a possibilidade de escolher, ou não, os termos a considerar nos processos de treino e de teste.

Os classificadores K-*Nearest Neighbour* e *Support Vector Machines* foram testados com três técnicas de selecção de termos, respectivamente, *Chi Square*, *Information Gain*, e finalmente, a técnica do Terceiro Momento em relação à média. Com o trabalho realizado é possível reter conclusões sobre o comportamento desta última técnica. O terceiro momento retorna resultados muito favoráveis, até mesmo superando as técnicas CHI e IG.

Da análise dos estudos realizados por outros autores (12) (17) (3) pode-se concluir que os resultados obtidos com a técnica do terceiro momento são equivalentes, e por vezes superando, os resultados obtidos por outras técnicas. Também é de notar que em muitos dos estudos, em que foi utilizado o *corpus* da *Reuters*-21578 (12) (17), é apenas tida em consideração a representação computacional por palavras. Sendo comum utilizar técnicas como listas de palavras não interessantes para remover termos que não contêm poder discriminante. No estudo descrito na

presente tese os termos foram seleccionados sem se utilizar qualquer técnica que acarrete dependências linguísticas.

O classificador RIPPER, sem utilizar técnicas de selecção de termos, consegue apresentar resultados com um desempenho próximo do SVM. Porém, o classificador RIPPER constrói as regras tendo em consideração todos os termos, verificando a co-ocorrência dos termos por cada documento da classe, e assim, criando regras para a classificação. Esta análise torna-se muito morosa, e pesada computacionalmente, quanto maior for o conjunto de termos, e o tamanho do *corpus*. Na experimentação realizada, o tempo que o sistema RIPPER demorou a construir o classificador, e a apresentar a classificação sobre o conjunto teste, foi no mínimo, sete vezes superior ao tempo que SVM demorou a construir o classificador, e a classificar o conjunto de teste, sobre o mesmo *corpus*. Para o classificador RIPPER os melhores resultados observaram-se para palavras, tendo a representação pelos primeiros 4 caracteres, o conjunto dos primeiros 4, 5 e 6 caracteres, e os pentagramas, apresentado resultados muito próximos.

Finalmente, este trabalho pretendeu ser um contributo para o aprofundamento do conhecimento sobre o impacto das técnicas de selecção de termos na performance dos classificadores. Comprovou-se que a técnica de selecção de termos 3M, com o classificador SVM, para as colecções R11 e R12, e relativamente às restantes técnicas de selecção de termos, obtêm um desempenho equivalente, e nalguns casos superior, como é o caso da representação pelos primeiros 4 caracteres.

Pode-se também concluir, que com outras representações de documentos, sem ser apenas por palavras, é possível obter desempenho melhor, ou equivalente, ao obtido com a representação por palavras. De destacar as representações computacionais por multi-palavras, sequências de 4 caracteres, e finalmente sequências de 456 caracteres.

Capítulo 5

Trabalho Futuro

No decorrer da experimentação, realizada nesta dissertação, observou-se que a pontuação, e reordenação, dos termos realizada por a ferramenta *Weka* para as técnicas *Information Gain* e *Chi-Square* existiam muitos termos que continham uma pontuação nula. A ordenação realizada no conjunto de termos, em que a ferramenta *Weka* indicou que o seu poder discriminante seria nulo, sugere ordem alfabética apesar de existirem alguns termos que surgem sem ser ordenados alfabeticamente. Seria interessante determinar qual seria a pontuação atribuída por uma outra ferramenta, ou implementar um programa em Java que realize essa tarefa. Assim, seria possível testar os classificadores com essa nova ordenação.

A ferramenta Weka possui uma outra limitação que poderá ter contribuído em uma diminuição da performance obtida com a técnica do Terceiro Momento. Os classificadores implementados nesta ferramenta, nomeadamente o *SVM* e *K-NN*, não têm em consideração o peso atribuído a cada um dos termos. Esta característica implica que a informação do peso atribuído a cada termo, por a técnica de selecção de termos, é ignorada pelo classificador. Em trabalho futuro seria interessante implementar os classificadores tendo em consideração o peso dos termos, ou utilizar outra ferramenta que tenha essa informação em consideração. Acredita-se que a técnica do Terceiro Momento apresentará melhores resultados sobre estas condições.

O desempenho obtido por Gonçalves e Quaresma (Gonçalves e Quaresma s.d.) sugere que em trabalho futuro após a selecção de termos com a técnica do terceiro momento estes sejam pesados por diferentes medidas de *term weighting*.

Em trabalho futuro seria interessante observar o comportamento da técnica do terceiro momento, com o *corpus* da *Reuters*-21578, utilizando o classificador, que foi descrito no capítulo 2.5.10, proposto por o Professor Joaquim F. Silva.

Nos resultados obtidos observou-se que existem classes que sugerem estarem co-relacionadas (por exemplo *corn* e *wheat*). Seria interessante observar a selecção de termos pela técnica do terceiro momento, e respectivo desempenho dos classificadores, unindo os documentos de classes que sugerem estar relacionadas. Esta sugestão de relacionamento é indicada pela matriz de confusão onde existem documentos classificados erradamente entre classes que se encontram tematicamente muito próximas.

Observou-se que com diferentes representações de documentos, além da usual representação por palavras, se obtêm resultados com um bom desempenho. Seria de interesse realizar um estudo tendo como representação de documentos todos os quadrigramas, pois com os primeiros 4 caracteres, globalmente, foram obtidos melhores resultados em relação às outras representações escolhidas.

Por fim seria interessante observar o comportamento dos classificadores, e técnicas de selecção de termos, utilizando como representação de documentos os primeiros 3 caracteres, e trigramas. Tamanho inferior a 3 considera-se que não terá interesse para categorização temática.

Apêndice

Nos apêndices A, B, C, D, E, F, G e H será possível consultar a informação do desempenho por classe com cada classificador, para cada representação de documentos.

No caso dos classificadores SVM (apêndice A e D) e KNN (apêndice B e E) encontra-se a informação do desempenho obtido tendo em consideração as três técnicas de selecção de termos, nomeadamente: Terceiro Momento, *Chi-Square*, e *Information Gain*.

Os desempenho por classe obtido com o classificador RIPPER para as sete representações de documentos encontram-se no apêndice A e F.

Os apêndices A, B, e C correspondem aos resultados obtidos com a colecção R11. E, os apêndices D, E e F correspondem aos resultados obtidos com a colecção R12.

Os apêndices G e H correspondem ao desempenho obtido com a colecção R4. O apêndice G corresponde à experimentação realizada com o classificador SVM para os melhores 4000 termos; enquanto o apêndice H corresponde à experimentação realizada com o classificador K-NN com o melhores 2000 termos.

Apêndice A

SVM: Resultados por classe com a colecção R11

A.1 Performance obtida com o classificador SVM utilizando a técnica do Terceiro Momento

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.784	0.968	0.866	0.956	livestock	0.667	0.667	0.667	0.991
alum	1	1	1	1	money-fx	0.55	0.458	0.5	0.904
bop	0.4	0.667	0.5	0.994	money-sup	0.4	0.4	0.4	0.941
carcass	0	0	0	0.902	nat-gas	0	0	0	0.989
cocoa	0	0	0	0.897	nzdlr	0	0	0	0.995
coffee	0.75	1	0.857	0.999	orange	1	0.333	0.5	0.997
copper	1	0.4	0.571	0.968	pet-chem	0	0	0	0.968
corn	0	0	0	0.983	platinum	0	0	0	0.411
cotton	1	1	1	1	potato	0	0	0	0.104
cpi	0.75	0.6	0.667	0.992	reserves	0.333	0.25	0.286	0.896
crude	0.667	0.609	0.636	0.909	rice	0	0	0	0.988
dlr	1	0.143	0.25	0.984	ship	0.615	0.571	0.593	0.881
earn	0.966	0.958	0.962	0.978	soybean	0	0	0	0.734
fuel	1	0.333	0.5	0.681	soy-meal	0	0	0	0.995
gas	0.5	1	0.667	0.999	soy-oil	0	0	0	0.801
gnp	0.5	0.667	0.571	0.994	strat-metal	0	0	0	0.611
gold	0.333	1	0.5	0.998	sugar	1	0.5	0.667	1
grain	0.333	0.333	0.333	0.732	tin	0	0	0	0.935
heat	0	0	0	1	trade	0.6	0.706	0.649	0.97
interest	0.515	0.68	0.586	0.966	veg-oil	0.5	0.333	0.4	0.937
ipi	1	0.2	0.333	0.982	wheat	0.75	0.75	0.75	0.988
iron-steel	0.667	0.5	0.571	0.97	wpi	1	0.5	0.667	1
jobs	0.5	0.5	0.5	0.983	yen	0	0	0	0.998
l-cattle	0	0	0	0.989	zinc	1	1	1	1
lei	0.5	0.5	0.5	0.997					

Tabela A.1.1: Performance por classe com representação de documentos por Palavras

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.679	0.911	0.778	0.905	lumber	0	0	0	0.973
alum	1	0.143	0.25	0.873	money-fx	0.762	0.516	0.615	0.936
barley	0	0	0	0.976	money-sup	0.667	0.286	0.4	0.714
bop	0.5	0.25	0.333	0.984	naphtha	0	0	0	0.5
carcass	0	0	0	0.709	nat-gas	1	0.125	0.222	0.803
cocoa	0.25	1	0.4	0.997	nkr	0	0	0	0.5
coffee	0.5	0.333	0.4	0.586	nzdlr	0	0	0	0.865
copper	0.333	0.25	0.286	0.889	orange	0.5	1	0.667	0.999
corn	0.5	0.4	0.444	0.964	pet-chem	0	0	0	0.617
cotton	0	0	0	0.668	potato	0	0	0	0.997
cpi	0	0	0	0.808	propane	0	0	0	0.5
crude	0.708	0.607	0.654	0.897	rapeseed	0	0	0	0.997
dlr	1	0.2	0.333	0.802	reserves	0	0	0	0.613
earn	0.934	0.939	0.936	0.962	rice	0	0	0	0.961
fuel	0	0	0	0.841	ship	0.375	0.462	0.414	0.951
gas	1	0.333	0.5	0.931	soybean	0	0	0	0.964
gnp	0.2	0.333	0.25	0.967	soy-oil	0	0	0	0.186
gold	0.667	0.333	0.444	0.847	strat-metal	1	0.333	0.5	0.649
grain	0	0	0	0.786	sugar	0.667	0.333	0.444	0.89
housing	0	0	0	0.398	trade	0.462	0.632	0.533	0.93
interest	0.429	0.632	0.511	0.946	veg-oil	1	0.5	0.667	0.901
ipi	0	0	0	0.979	wheat	0.231	0.375	0.286	0.857
iron-steel	0.333	0.5	0.4	0.993	wpi	0.667	0.4	0.5	0.994
jobs	0	0	0	0.994	yen	0	0	0	1
livestock	0	0	0	0.319					

Tabela A.1.2: Performance por classe com representação de documentos por Multi-palavras

A.1 Performance obtida com o classificador SVM utilizando a técnica do Terceiro Momento

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.824	0.911	0.865	0.939	lumber	0	0	0	0.908
alum	1	0.429	0.6	0.957	money-fx	0.727	0.516	0.604	0.953
barley	0	0	0	0.994	money-sup	0.667	0.571	0.615	0.957
bop	0.667	0.5	0.571	0.979	naphtha	0	0	0	0.5
carcass	0	0	0	0.986	nat-gas	0	0	0	0.906
cocoa	1	1	1	1	nkr	0	0	0	0.5
coffee	1	0.5	0.667	0.953	nzdlr	0	0	0	0.997
copper	1	0.5	0.667	0.934	orange	1	1	1	1
corn	1	0.4	0.571	0.985	pet-chem	0	0	0	0.821
cotton	1	0.667	0.8	1	potato	1	1	1	1
cpi	0.5	0.25	0.333	0.969	propane	0	0	0	0.5
crude	0.606	0.714	0.656	0.958	rapeseed	1	1	1	1
dlr	0.667	0.4	0.5	0.982	reserves	1	0.25	0.4	0.779
earn	0.844	0.967	0.901	0.946	rice	1	0.333	0.5	0.991
fuel	0	0	0	0.997	ship	0.438	0.538	0.483	0.984
gas	0.75	1	0.857	0.999	soybean	0.333	0.5	0.4	0.997
gnp	0.5	0.667	0.571	0.988	soy-oil	0	0	0	0.223
gold	0.6	0.5	0.545	0.992	strat-metal	1	0.333	0.5	0.507
grain	0	0	0	0.912	sugar	1	0.333	0.5	0.857
housing	0	0	0	0.63	trade	0.667	0.737	0.7	0.947
interest	0.583	0.737	0.651	0.906	veg-oil	0.5	0.5	0.5	0.979
ipi	0	0	0	0.993	wheat	0.889	1	0.941	0.999
iron-steel	0.5	0.5	0.5	0.958	wpi	1	0.6	0.75	0.998
jobs	0.333	1	0.5	0.998	yen	0	0	0	0.997
livestock	1	0.5	0.667	0.93					

Tabela A.1.3: Performance por classe com representação de documentos por sequências dos primeiros 4 caracteres

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.838	0.832	0.835	0.916	jet	0	0	0	0.5
alum	1	0.75	0.857	0.999	jobs	0.667	0.667	0.667	0.954
barley	0	0	0	1	lead	1	1	1	1
bop	0	0	0	0.981	livestock	1	1	1	1
carcass	0	0	0	0.974	money-fx	0.526	0.5	0.513	0.908
cocoa	1	0.5	0.667	0.755	money-sup	0.75	0.5	0.6	0.841
coffee	1	0.4	0.571	0.998	nat-gas	0.25	0.333	0.286	0.947
copper	1	1	1	1	pet-chem	0	0	0	0.75
corn	0.333	0.333	0.333	0.975	potato	0	0	0	0.998
cotton	1	0.5	0.667	1	reserves	1	1	1	1
cpi	0.5	0.5	0.5	0.965	rice	1	1	1	1
crude	0.72	0.72	0.72	0.943	ship	0.8	0.471	0.593	0.927
dlr	1	0.5	0.667	0.993	soybean	0	0	0	0.989
earn	0.793	0.964	0.87	0.915	soy-meal	0	0	0	0.953
fuel	0	0	0	0.752	strat-metal	0	0	0	0.385
gas	0	0	0	0.939	sugar	1	0.2	0.333	0.981
gnp	0.667	0.4	0.5	0.99	tea	0	0	0	0.987
gold	0	0	0	0.989	trade	0.75	0.8	0.774	0.984
grain	0	0	0	0.978	veg-oil	0	0	0	0.911
groundnut	0	0	0	0.5	wheat	0.5	0.8	0.615	0.995
heat	1	1	1	1	wpi	1	1	1	1
instal-debt	0	0	0	0.5	yen	0	0	0	0.995
interest	0.588	0.476	0.526	0.884	zinc	0.5	1	0.667	0.999
ipi	1	1	1	1					
iron-steel	0.667	0.5	0.571	0.76					

Tabela A.1.4: Performance por classe com representação de documentos por sequências dos primeiros 5 caracteres

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.855	0.778	0.815	0.905	jobs	1	0.333	0.5	0.946
alum	0.8	1	0.889	0.999	lead	0	0	0	0.702
barley	1	1	1	1	lei	0	0	0	1
bop	0.2	0.2	0.2	0.972	livestock	1	0.5	0.667	0.923
carcass	1	0.333	0.5	0.986	meal-feed	0	0	0	0.5
cocoa	0	0	0	0.992	money-fx	0.789	0.556	0.652	0.878
coffee	1	0.2	0.333	0.992	money-sup	1	0.2	0.333	0.894
copper	1	1	1	1	nat-gas	0	0	0	0.967
corn	0.5	0.2	0.286	0.773	nkr	0.5	1	0.667	0.999
cotton	1	1	1	1	orange	0	0	0	1
cpi	1	0.25	0.4	0.996	palm-oil	0	0	0	0.39
crude	0.64	0.552	0.593	0.933	pet-chem	0	0	0	0.724
dlr	1	0.333	0.5	0.881	rapeseed	1	1	1	1
dmk	0	0	0	0.5	reserves	0.5	0.5	0.5	0.982
earn	0.707	0.971	0.818	0.885	rice	0	0	0	0.976
fuel	1	1	1	1	ship	0.583	0.389	0.467	0.863
gnp	0.5	0.5	0.5	0.98	soybean	0	0	0	1
gold	1	0.286	0.444	0.996	sugar	0.5	0.667	0.571	0.95
grain	0	0	0	0.837	tin	0	0	0	0.01
groundnut	0	0	0	0.428	trade	0.429	0.643	0.514	0.917
heat	0	0	0	0.026	veg-oil	1	0.5	0.667	0.958
income	0	0	0	0.999	wheat	0.625	0.714	0.667	0.91
interest	0.5	0.304	0.378	0.778	wpi	1	0.667	0.8	1
iron-steel	0	0	0	0.913	zinc	1	0.333	0.5	0.545
jet	0	0	0	0.5					

Tabela A.1.5: Performance por classe com representação de documentos por sequências dos primeiros 6 caracteres

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.743	0.846	0.791	0.904	lumber	0	0	0	0.951
alum	1	0.714	0.833	0.991	money-fx	0.696	0.516	0.593	0.954
barley	1	0.5	0.667	0.999	money-sup	0.8	0.571	0.667	0.886
bop	0.4	0.5	0.444	0.983	naphtha	0	0	0	0.5
carcass	0	0	0	0.957	nat-gas	0	0	0	0.753
cocoa	1	1	1	1	nkr	0	0	0	0.5
coffee	1	0.5	0.667	0.98	nzdlr	0	0	0	0.975
copper	1	0.75	0.857	0.988	orange	1	1	1	1
corn	1	0.2	0.333	0.818	pet-chem	0	0	0	0.868
cotton	1	0.667	0.8	1	potato	0	0	0	0.999
cpi	0.5	0.25	0.333	0.884	propane	0	0	0	0.5
crude	0.679	0.679	0.679	0.942	rapeseed	0	0	0	1
dlr	0.667	0.4	0.5	0.984	reserves	0	0	0	0.746
earn	0.83	0.967	0.893	0.943	rice	1	0.333	0.5	0.969
fuel	0	0	0	0.873	ship	0.429	0.462	0.444	0.923
gas	0.6	1	0.75	0.998	soybean	0.5	0.5	0.5	0.999
gnp	0.5	0.667	0.571	0.994	soy-oil	0	0	0	0.958
gold	0.5	0.167	0.25	0.736	strat-metal	1	0.333	0.5	0.866
grain	0	0	0	0.925	sugar	0.4	0.333	0.364	0.644
housing	0	0	0	0.976	trade	0.6	0.632	0.615	0.966
interest	0.5	0.632	0.558	0.909	veg-oil	0	0	0	0.88
ipi	1	0.5	0.667	0.996	wheat	0.625	0.625	0.625	0.991
iron-steel	0.5	0.5	0.5	0.997	wpi	1	0.4	0.571	0.974
jobs	0.5	1	0.667	0.999	yen	0	0	0	0.996
livestock	1	0.5	0.667	0.776					

Tabela A.1.6: Performance por classe com representação de documentos por sequências dos primeiros 4, 5 e 6 caracteres

A.1 Performance obtida com o classificador SVM utilizando a técnica do Terceiro Momento

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.771	0.745	0.758	0.874	jet	0	0	0	0.5
alum	1	0.5	0.667	0.997	jobs	0.333	0.333	0.333	0.953
barley	0	0	0	1	lead	1	0.5	0.667	1
bop	0.333	0.25	0.286	0.977	livestock	1	1	1	1
carcass	1	0.5	0.667	0.937	money-fx	0.688	0.55	0.611	0.911
cocoa	0.5	0.5	0.5	0.866	money-sup	1	0.5	0.667	0.887
coffee	1	0.4	0.571	0.995	nat-gas	0	0	0	0.937
copper	0.333	1	0.5	0.998	pet-chem	0	0	0	0.863
corn	0	0	0	0.653	potato	0	0	0	0.007
cotton	1	1	1	1	reserves	1	1	1	1
cpi	0.333	0.5	0.4	0.982	rice	1	1	1	1
crude	0.636	0.56	0.596	0.882	ship	0.769	0.588	0.667	0.849
dlr	1	0.5	0.667	0.988	soybean	0	0	0	0.998
earn	0.769	0.959	0.854	0.909	soy-meal	0	0	0	0.989
fuel	0	0	0	0.814	strat-metal	0	0	0	0.136
gas	0	0	0	0.798	sugar	0.5	0.2	0.286	0.6
gnp	0.5	0.2	0.286	0.969	tea	0	0	0	1
gold	0	0	0	0.988	trade	0.714	0.667	0.69	0.992
grain	0	0	0	0.954	veg-oil	0	0	0	0.945
groundnut	0	0	0	0.5	wheat	0.5	0.8	0.615	0.971
heat	1	1	1	1	wpi	1	1	1	1
instal-debt	0	0	0	0.5	yen	0	0	0	0.996
interest	0.722	0.619	0.667	0.928	zinc	0	0	0	0.946
ipi	1	1	1	1					
iron-steel	0.667	0.5	0.571	0.771					

Tabela A.1.7: Performance por classe com representação de documentos por Pentagramas

A.2 Performance obtida com o classificador SVM utilizando a técnica *Chi-Square*

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.667	0.984	0.795	0.924	livestock	0	0	0	0.313
alum	0.667	0.5	0.571	0.889	money-fx	0.682	0.625	0.652	0.921
bop	1	0.333	0.5	0.794	money-sup	0.333	0.4	0.364	0.961
carcass	0	0	0	0.971	nat-gas	0.5	0.5	0.5	0.999
cocoa	0	0	0	0.897	nzdlr	0	0	0	1
coffee	1	0.667	0.8	0.997	orange	1	0.333	0.5	0.983
copper	1	0.2	0.333	0.965	pet-chem	0	0	0	0.818
corn	0	0	0	0.982	platinum	0	0	0	0.454
cotton	0	0	0	0.901	potato	0	0	0	0.062
cpi	0.5	0.4	0.444	0.989	reserves	0.5	0.25	0.333	0.88
crude	0.737	0.609	0.667	0.936	rice	0	0	0	0.924
dlr	0	0	0	0.976	ship	0.75	0.643	0.692	0.936
earn	0.987	0.933	0.959	0.977	soybean	0	0	0	0.588
fuel	1	0.333	0.5	0.992	soy-meal	0	0	0	0.976
gas	1	1	1	1	soy-oil	0	0	0	0.862
gnp	0.5	0.333	0.4	0.981	strat-metal	0	0	0	0.689
gold	1	1	1	1	sugar	0	0	0	0.993
grain	0	0	0	0.825	tin	0.2	0.25	0.222	0.898
heat	0	0	0	0.912	trade	0.583	0.824	0.683	0.989
interest	0.6	0.6	0.6	0.887	veg-oil	1	0.333	0.5	0.959
ipi	1	0.2	0.333	0.96	wheat	0.857	0.75	0.8	0.99
iron-steel	0	0	0	0.878	wpi	1	1	1	1
jobs	0	0	0	0.959	yen	0	0	0	0.987
l-cattle	0	0	0	0.963	zinc	0	0	0	0.771
lei	0	0	0	0.677					

Tabela A.2.1: Performance por classe com representação de documentos por Palavras

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.716	0.939	0.813	0.918	lei	0	0	0	0.987
alum	0.5	0.667	0.571	0.94	livestock	0	0	0	0.591
barley	0	0	0	0.941	meal-feed	0	0	0	0.5
bop	0.5	0.5	0.5	0.997	money-fx	0.579	0.458	0.512	0.88
carcass	1	1	1	1	money-sup	1	0.625	0.769	0.87
cocoa	0	0	0	0.581	nat-gas	0.4	0.25	0.308	0.809
coffee	1	0.143	0.25	0.816	nzdlr	0	0	0	0.5
copper	0	0	0	0.325	orange	1	0.333	0.5	0.818
corn	0	0	0	0.703	pet-chem	0	0	0	0.915
cotton	0	0	0	0.354	platinum	0	0	0	0.936
cpi	0	0	0	0.859	potato	0	0	0	0.036
crude	0.704	0.704	0.704	0.921	propane	0	0	0	0.5
dlr	0	0	0	0.812	reserves	1	0.333	0.5	0.864
earn	0.948	0.971	0.959	0.977	ship	0.615	0.571	0.593	0.957
gas	0.5	0.5	0.5	0.68	soybean	0	0	0	0.999
gnp	0	0	0	0.992	soy-meal	0	0	0	0.539
gold	0.333	0.25	0.286	0.967	sugar	1	0.286	0.444	0.788
grain	0.333	0.333	0.333	0.711	tin	0	0	0	0.778
heat	0	0	0	0.977	trade	0.571	0.706	0.632	0.986
income	0	0	0	0.98	veg-oil	0	0	0	0.817
interest	0.519	0.737	0.609	0.925	wheat	0.714	0.714	0.714	0.997
ipi	0	0	0	0.627	wpi	0	0	0	0.85
iron-steel	1	0.5	0.667	0.666	yen	1	1	1	1
jobs	1	1	1	1					
lead	0	0	0	0.763					

Tabela A.2.2: Performance por classe com representação de documentos por Multi-palavras

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.786	0.894	0.837	0.933	lumber	0	0	0	0.955
alum	1	0.571	0.727	0.981	money-fx	0.769	0.323	0.455	0.843
barley	0	0	0	0.997	money-sup	0.5	0.429	0.462	0.904
bop	0.4	0.5	0.444	0.975	naphtha	0	0	0	0.5
carcass	0	0	0	0.974	nat-gas	0.5	0.125	0.2	0.889
cocoa	0	0	0	1	nkr	0	0	0	0.5
coffee	1	0.333	0.5	0.93	nzdlr	0	0	0	0.992
copper	1	0.25	0.4	0.987	orange	1	1	1	1
corn	0	0	0	0.816	pet-chem	0	0	0	0.836
cotton	1	0.667	0.8	0.992	potato	1	1	1	1
cpi	0.333	0.25	0.286	0.971	propane	0	0	0	0.5
crude	0.708	0.607	0.654	0.939	rapeseed	0	0	0	0.988
dlr	1	0.2	0.333	0.808	reserves	1	0.25	0.4	0.803
earn	0.739	0.948	0.831	0.892	rice	0	0	0	0.919
fuel	0	0	0	0.941	ship	0.467	0.538	0.5	0.952
gas	0.667	0.667	0.667	0.995	soybean	0	0	0	0.942
gnp	0.5	0.333	0.4	0.989	soy-oil	0	0	0	0.536
gold	0.5	0.333	0.4	0.985	strat-metal	1	0.333	0.5	0.5
grain	0	0	0	0.908	sugar	0.8	0.667	0.727	0.856
housing	0	0	0	0.903	trade	0.5	0.526	0.513	0.962
interest	0.429	0.474	0.45	0.849	veg-oil	0	0	0	0.932
ipi	0	0	0	0.99	wheat	0.889	1	0.941	0.999
iron-steel	0.5	0.5	0.5	0.95	wpi	1	0.2	0.333	0.989
jobs	1	1	1	1	yen	0	0	0	1
livestock	0	0	0	0.805					

Tabela A.2.3: Performance por classe com representação de documentos por sequências dos primeiros 4 caracteres

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.864	0.852	0.858	0.922	jet	0	0	0	0.5
alum	1	1	1	1	jobs	1	0.333	0.5	0.928
barley	0	0	0	1	lead	0	0	0	0.985
bop	0.333	0.25	0.286	0.837	livestock	0	0	0	0.475
carcass	0	0	0	0.982	money-fx	0.684	0.65	0.667	0.942
cocoa	0.5	0.5	0.5	0.915	money-sup	0.5	0.167	0.25	0.855
coffee	1	0.4	0.571	0.992	nat-gas	0	0	0	0.781
copper	1	1	1	1	pet-chem	0	0	0	0.67
corn	0.125	0.333	0.182	0.877	potato	0	0	0	0.031
cotton	0.667	1	0.8	0.999	reserves	1	0.333	0.5	0.997
cpi	0.5	0.5	0.5	0.583	rice	0	0	0	0.993
crude	0.591	0.52	0.553	0.894	ship	0.5	0.294	0.37	0.833
dlr	0.333	0.5	0.4	0.992	soybean	0	0	0	0.763
earn	0.727	0.973	0.832	0.876	soy-meal	0	0	0	0.529
fuel	0	0	0	0.509	strat-metal	0	0	0	0.63
gas	0	0	0	0.901	sugar	1	0.2	0.333	0.953
gnp	0.75	0.6	0.667	0.988	tea	0	0	0	0.991
gold	0	0	0	0.942	trade	0.7	0.467	0.56	0.95
grain	0	0	0	0.979	veg-oil	0	0	0	0.964
groundnut	0	0	0	0.5	wheat	0.625	1	0.769	0.997
heat	1	1	1	1	wpi	0	0	0	0.982
instal-debt	0	0	0	0.5	yen	0	0	0	1
interest	0.818	0.429	0.563	0.768	zinc	0	0	0	1
ipi	0	0	0	0.989					
iron-steel	1	0.25	0.4	0.791					

Tabela A.2.4: Performance por classe com representação de documentos por sequências dos primeiros 5 caracteres

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.835	0.771	0.801	0.894	jobs	1	0.667	0.8	0.994
alum	0.8	1	0.889	0.999	lead	0	0	0	0.693
barley	1	1	1	1	lei	0	0	0	1
bop	0.25	0.2	0.222	0.901	livestock	0	0	0	0.601
carcass	1	0.667	0.8	0.956	meal-feed	0	0	0	0.5
cocoa	1	1	1	0.998	money-fx	0.625	0.37	0.465	0.856
coffee	1	0.2	0.333	0.969	money-sup	0	0	0	0.827
copper	1	1	1	1	nat-gas	1	0.5	0.667	0.993
corn	0	0	0	0.923	nkr	0	0	0	0.604
cotton	0	0	0	0.809	orange	1	1	1	1
cpi	0	0	0	0.872	palm-oil	0	0	0	0.574
crude	0.667	0.483	0.56	0.898	pet-chem	0	0	0	0.996
dlr	1	0.5	0.667	0.837	rapeseed	0	0	0	0.998
dmk	0	0	0	0.5	reserves	0.5	0.5	0.5	0.985
earn	0.653	0.971	0.781	0.855	rice	0	0	0	0.962
fuel	0	0	0	0.998	ship	1	0.278	0.435	0.864
gnp	0.333	0.25	0.286	0.981	soybean	0	0	0	0.955
gold	0.5	0.286	0.364	0.985	sugar	1	0.333	0.5	0.907
grain	1	0.333	0.5	0.828	tin	0	0	0	0.026
groundnut	0	0	0	0.491	trade	0.563	0.643	0.6	0.961
heat	0	0	0	0.5	veg-oil	1	0.25	0.4	0.946
income	0	0	0	0.968	wheat	0.636	1	0.778	0.996
interest	0.467	0.304	0.368	0.74	wpi	1	0.333	0.5	0.992
iron-steel	0	0	0	0.953	zinc	0	0	0	0.289
jet	0	0	0	0.5					

Tabela A.2.5: Performance por classe com representação de documentos por sequências dos primeiros 6 caracteres

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.824	0.878	0.85	0.931	lumber	1	1	1	1
alum	1	0.714	0.833	0.99	money-fx	0.833	0.484	0.612	0.93
barley	0	0	0	0.966	money-sup	0	0	0	0.821
bop	0.5	0.25	0.333	0.983	naphtha	0	0	0	0.5
carcass	0	0	0	0.932	nat-gas	0	0	0	0.854
cocoa	0.5	1	0.667	0.999	nkr	0	0	0	0.5
coffee	1	0.5	0.667	0.975	nzdlr	0	0	0	0.993
copper	0.5	0.25	0.333	0.992	orange	1	1	1	1
corn	0	0	0	0.724	pet-chem	0	0	0	0.537
cotton	0	0	0	0.954	potato	0	0	0	0.92
cpi	0.6	0.75	0.667	0.996	propane	0	0	0	0.5
crude	0.606	0.714	0.656	0.969	rapeseed	0	0	0	0.981
dlr	0	0	0	0.786	reserves	0.5	0.5	0.5	0.806
earn	0.735	0.953	0.83	0.89	rice	0	0	0	0.989
fuel	0	0	0	0.994	ship	0.667	0.462	0.545	0.986
gas	0.75	1	0.857	0.999	soybean	0	0	0	0.961
gnp	0.333	0.333	0.333	0.988	soy-oil	0	0	0	0.933
gold	0.75	0.5	0.6	0.99	strat-metal	1	0.333	0.5	0.757
grain	0	0	0	0.968	sugar	0.8	0.667	0.727	0.993
housing	0	0	0	0.706	trade	0.667	0.842	0.744	0.973
interest	0.524	0.579	0.55	0.894	veg-oil	0	0	0	0.955
ipi	0	0	0	0.954	wheat	0.875	0.875	0.875	0.999
iron-steel	1	0.5	0.667	0.977	wpi	0	0	0	0.996
jobs	0	0	0	0.996	yen	0	0	0	0.996
livestock	0	0	0	0.538					

Tabela A.2.7: Performance por classe com representação de documentos por sequências dos primeiros 4, 5 e 6 caracteres

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.896	0.866	0.881	0.937	jobs	0	0	0	0.95
alum	1	1	1	1	lead	0	0	0	0.855
barley	0	0	0	0.952	livestock	0	0	0	0.272
bop	0.333	0.25	0.286	0.92	money-fx	0.737	0.7	0.718	0.979
carcass	1	0.5	0.667	0.963	money-sup	1	0.167	0.286	0.918
cocoa	0.333	0.5	0.4	0.879	nat-gas	0	0	0	0.732
coffee	0.6	0.6	0.6	0.996	pet-chem	0	0	0	0.737
copper	1	1	1	1	potato	0	0	0	0.017
corn	0	0	0	0.972	reserves	1	1	1	1
cotton	0	0	0	0.967	rice	0	0	0	0.997
cpi	0.5	0.5	0.5	0.607	ship	0.556	0.294	0.385	0.904
crude	0.778	0.56	0.651	0.947	soybean	0	0	0	0.942
dlr	0	0	0	0.984	soy-meal	0	0	0	0.936
fuel	0	0	0	0.88	strat-metal	0	0	0	0.572
gas	0	0	0	0.815	sugar	1	0.2	0.333	0.975
gnp	0.667	0.4	0.5	0.99	tea	0	0	0	0.983
gold	0	0	0	0.971	trade	0.75	0.6	0.667	0.984
grain	0.5	0.25	0.333	0.968	veg-oil	0.5	0.333	0.4	0.959
groundnut	0	0	0	0.5	wheat	0.625	1	0.769	0.997
heat	0.5	1	0.667	0.999	wpi	0	0	0	0.975
instal-debt	0	0	0	0.5	yen	0	0	0	0.996
interest	0.667	0.571	0.615	0.812	zinc	0	0	0	0.569
ipi	0	0	0	0.962					
iron-steel	1	0.25	0.4	0.507					
jet	0	0	0	0.5					

Tabela A.2.7: Performance por classe com representação de documentos por Pentagramas

A.3 Performance obtida com o classificador SVM utilizando a técnica *Information Gain*

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.667	0.984	0.795	0.924	livestock	0	0	0	0.195
alum	0.667	0.5	0.571	0.888	money-fx	0.682	0.625	0.652	0.921
bop	1	0.333	0.5	0.791	money-sup	0.333	0.4	0.364	0.961
carcass	0	0	0	0.975	nat-gas	0.5	0.5	0.5	0.999
cocoa	0	0	0	0.895	nzdlr	0	0	0	1
coffee	1	0.667	0.8	0.997	orange	1	0.333	0.5	0.983
copper	1	0.2	0.333	0.965	pet-chem	0	0	0	0.812
corn	0	0	0	0.982	platinum	0	0	0	0.454
cotton	0	0	0	0.897	potato	0	0	0	0.062
cpi	0.5	0.4	0.444	0.989	reserves	0.5	0.25	0.333	0.881
crude	0.737	0.609	0.667	0.936	rice	0	0	0	0.925
dlr	0	0	0	0.976	ship	0.75	0.643	0.692	0.936
earn	0.987	0.933	0.959	0.977	soybean	0	0	0	0.538
fuel	1	0.333	0.5	0.992	soy-meal	0	0	0	0.976
gas	1	1	1	1	soy-oil	0	0	0	0.862
gnp	0.5	0.333	0.4	0.98	strat-metal	0	0	0	0.692
gold	1	1	1	1	sugar	0	0	0	0.993
grain	0	0	0	0.827	tin	0.2	0.25	0.222	0.898
heat	0	0	0	0.913	trade	0.583	0.824	0.683	0.989
interest	0.6	0.6	0.6	0.887	veg-oil	1	0.333	0.5	0.96
ipi	1	0.2	0.333	0.929	wheat	0.857	0.75	0.8	0.99
iron-steel	0	0	0	0.854	wpi	1	1	1	1
jobs	0	0	0	0.948	yen	0	0	0	0.987
l-cattle	0	0	0	0.963	zinc	0	0	0	0.742
lei	0	0	0	0.675					

Tabela A.3.1: Performance por classe com representação de documentos por Palavras

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.716	0.939	0.813	0.918	lei	0	0	0	0.987
alum	0.5	0.667	0.571	0.94	livestock	0	0	0	0.591
barley	0	0	0	0.941	meal-feed	0	0	0	0.5
bop	0.5	0.5	0.5	0.997	money-fx	0.579	0.458	0.512	0.88
carcass	1	1	1	1	money-sup	1	0.625	0.769	0.87
cocoa	0	0	0	0.581	nat-gas	0.4	0.25	0.308	0.809
coffee	1	0.143	0.25	0.816	nzdlr	0	0	0	0.5
copper	0	0	0	0.325	orange	1	0.333	0.5	0.818
corn	0	0	0	0.703	pet-chem	0	0	0	0.915
cotton	0	0	0	0.354	platinum	0	0	0	0.936
cpi	0	0	0	0.859	potato	0	0	0	0.036
crude	0.704	0.704	0.704	0.921	propane	0	0	0	0.5
dlr	0	0	0	0.812	reserves	1	0.333	0.5	0.864
earn	0.948	0.971	0.959	0.977	ship	0.615	0.571	0.593	0.957
gas	0.5	0.5	0.5	0.68	soybean	0	0	0	0.999
gnp	0	0	0	0.992	soy-meal	0	0	0	0.539
gold	0.333	0.25	0.286	0.967	sugar	1	0.286	0.444	0.788
grain	0.333	0.333	0.333	0.711	tin	0	0	0	0.778
heat	0	0	0	0.977	trade	0.571	0.706	0.632	0.986
income	0	0	0	0.98	veg-oil	0	0	0	0.817
interest	0.519	0.737	0.609	0.925	wheat	0.714	0.714	0.714	0.997
ipi	0	0	0	0.627	wpi	0	0	0	0.85
iron-steel	1	0.5	0.667	0.666	yen	1	1	1	1
jobs	1	1	1	1					
lead	0	0	0	0.763					

Tabela A.3.2: Performance por classe com representação de documentos por Multi-palavras

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.822	0.902	0.86	0.94	lumber	0	0	0	0.963
alum	1	0.571	0.727	0.967	money-fx	0.769	0.323	0.455	0.847
barley	0	0	0	0.997	money-sup	0.5	0.429	0.462	0.865
bop	0.4	0.5	0.444	0.974	naphtha	0	0	0	0.5
carcass	0	0	0	0.976	nat-gas	0.333	0.125	0.182	0.891
cocoa	0	0	0	1	nkr	0	0	0	0.5
coffee	1	0.333	0.5	0.93	nzdlr	0	0	0	0.99
copper	1	0.25	0.4	0.987	orange	1	1	1	1
corn	0	0	0	0.864	pet-chem	0	0	0	0.893
cotton	1	0.667	0.8	0.993	potato	1	1	1	1
cpi	0.333	0.25	0.286	0.972	propane	0	0	0	0.5
crude	0.72	0.643	0.679	0.94	rapeseed	0	0	0	0.988
dlr	1	0.2	0.333	0.818	reserves	0	0	0	0.758
earn	0.739	0.948	0.831	0.891	rice	0	0	0	0.925
fuel	0	0	0	0.946	ship	0.467	0.538	0.5	0.935
gas	0.667	0.667	0.667	0.995	soybean	0	0	0	0.944
gnp	0.5	0.333	0.4	0.988	soy-oil	0	0	0	0.578
gold	0.5	0.333	0.4	0.986	strat-metal	1	0.333	0.5	0.624
grain	0	0	0	0.924	sugar	0.8	0.667	0.727	0.91
housing	0	0	0	0.893	trade	0.524	0.579	0.55	0.963
interest	0.417	0.526	0.465	0.85	veg-oil	0	0	0	0.934
ipi	0	0	0	0.99	wheat	0.889	1	0.941	0.999
iron-steel	0.5	0.5	0.5	0.647	wpi	1	0.2	0.333	0.99
jobs	1	1	1	1	yen	0	0	0	1
livestock	1	0.5	0.667	0.769	acq	0.822	0.902	0.86	0.94

Tabela A.3.3: Performance por classe com representação de documentos por sequências dos primeiros 4 caracteres

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.864	0.852	0.858	0.922	jet	0	0	0	0.5
alum	1	1	1	1	jobs	1	0.333	0.5	0.928
barley	0	0	0	1	lead	0	0	0	0.985
bop	0.333	0.25	0.286	0.837	livestock	0	0	0	0.426
carcass	0	0	0	0.982	money-fx	0.684	0.65	0.667	0.942
cocoa	0.5	0.5	0.5	0.934	money-sup	0	0	0	0.854
coffee	1	0.4	0.571	0.991	nat-gas	0	0	0	0.808
copper	1	1	1	1	pet-chem	0	0	0	0.67
corn	0.125	0.333	0.182	0.877	potato	0	0	0	0.031
cotton	0.667	1	0.8	0.999	reserves	1	0.333	0.5	0.997
cpi	0.5	0.5	0.5	0.584	rice	0	0	0	0.993
crude	0.591	0.52	0.553	0.894	ship	0.5	0.294	0.37	0.833
dlr	0.333	0.5	0.4	0.992	soybean	0	0	0	0.763
earn	0.725	0.973	0.831	0.874	soy-meal	0	0	0	0.529
fuel	0	0	0	0.509	strat-metal	0	0	0	0.629
gas	0	0	0	0.917	sugar	1	0.2	0.333	0.952
gnp	0.75	0.6	0.667	0.988	tea	0	0	0	0.991
gold	0	0	0	0.941	trade	0.7	0.467	0.56	0.951
grain	0	0	0	0.979	veg-oil	0	0	0	0.964
groundnut	0	0	0	0.5	wheat	0.625	1	0.769	0.997
heat	1	1	1	1	wpi	0	0	0	0.982
instal-debt	0	0	0	0.5	yen	0	0	0	1
interest	0.818	0.429	0.563	0.768	zinc	0	0	0	1
ipi	0	0	0	0.989					
iron-steel	1	0.25	0.4	0.79					

Tabela A.3.4: Performance por classe com representação de documentos por sequências dos primeiros 5 caracteres

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.847	0.771	0.807	0.893	jobs	1	0.667	0.8	0.994
alum	0.8	1	0.889	0.999	lead	0	0	0	0.693
barley	1	1	1	1	lei	0	0	0	1
bop	0.25	0.2	0.222	0.901	livestock	0	0	0	0.6
carcass	1	0.667	0.8	0.956	meal-feed	0	0	0	0.5
cocoa	0.333	1	0.5	0.996	money-fx	0.625	0.37	0.465	0.856
coffee	1	0.2	0.333	0.969	money-sup	0	0	0	0.829
copper	1	1	1	1	nat-gas	1	0.5	0.667	0.993
corn	0	0	0	0.923	nkr	0	0	0	0.604
cotton	0	0	0	0.808	orange	1	1	1	1
cpi	0	0	0	0.871	palm-oil	0	0	0	0.574
crude	0.667	0.483	0.56	0.899	pet-chem	0	0	0	0.996
dlr	1	0.5	0.667	0.837	rapeseed	0	0	0	0.998
dmk	0	0	0	0.5	reserves	0.5	0.5	0.5	0.985
earn	0.653	0.971	0.781	0.855	rice	0	0	0	0.962
fuel	0	0	0	0.998	ship	1	0.278	0.435	0.863
gnp	0.333	0.25	0.286	0.981	soybean	0	0	0	0.956
gold	0.5	0.286	0.364	0.984	sugar	1	0.333	0.5	0.879
grain	1	0.333	0.5	0.83	tin	0	0	0	0.026
groundnut	0	0	0	0.491	trade	0.563	0.643	0.6	0.96
heat	0	0	0	0.5	veg-oil	1	0.25	0.4	0.946
income	0	0	0	0.968	wheat	0.636	1	0.778	0.996
interest	0.467	0.304	0.368	0.741	wpi	1	0.333	0.5	0.992
iron-steel	0	0	0	0.954	zinc	0	0	0	0.289
jet	0	0	0	0.5					

Tabela A.3.5: Performance por classe com representação de documentos por sequências dos primeiros 6 caracteres

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.831	0.878	0.854	0.931	lumber	1	1	1	1
alum	1	0.714	0.833	0.991	money-fx	0.833	0.484	0.612	0.936
barley	0	0	0	0.966	money-sup	0	0	0	0.82
bop	0.5	0.25	0.333	0.983	naphtha	0	0	0	0.5
carcass	0	0	0	0.932	nat-gas	0	0	0	0.853
cocoa	0.5	1	0.667	0.999	nkr	0	0	0	0.5
coffee	1	0.5	0.667	0.969	nzdlr	0	0	0	0.993
copper	0.5	0.25	0.333	0.992	orange	1	1	1	1
corn	0	0	0	0.707	pet-chem	0	0	0	0.548
cotton	0	0	0	0.949	potato	0	0	0	0.92
cpi	0.6	0.75	0.667	0.996	propane	0	0	0	0.5
crude	0.606	0.714	0.656	0.969	rapeseed	0	0	0	0.981
dlr	0	0	0	0.786	reserves	0.5	0.5	0.5	0.805
earn	0.735	0.953	0.83	0.889	rice	0	0	0	0.99
fuel	0	0	0	0.994	ship	0.667	0.462	0.545	0.986
gas	0.75	1	0.857	0.999	soybean	0	0	0	0.961
gnp	0.333	0.333	0.333	0.988	soy-oil	0	0	0	0.933
gold	0.75	0.5	0.6	0.99	strat-metal	1	0.333	0.5	0.758
grain	0	0	0	0.968	sugar	0.8	0.667	0.727	0.993
housing	0	0	0	0.427	trade	0.667	0.842	0.744	0.973
interest	0.524	0.579	0.55	0.894	veg-oil	0	0	0	0.955
ipi	0	0	0	0.95	wheat	0.875	0.875	0.875	0.999
iron-steel	1	0.5	0.667	0.977	wpi	0	0	0	0.996
jobs	0	0	0	0.997	yen	0	0	0	0.996
livestock	0	0	0	0.391					

Tabela A.3.6: Performance por classe com representação de documentos por sequências dos primeiros 4, 5 e 6 caracteres

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.896	0.866	0.881	0.937	jet	0	0	0	0.5
alum	1	1	1	1	jobs	0	0	0	0.947
barley	0	0	0	0.952	lead	0	0	0	0.857
bop	0.333	0.25	0.286	0.921	livestock	0	0	0	0.257
carcass	1	0.5	0.667	0.97	money-fx	0.737	0.7	0.718	0.979
cocoa	0.333	0.5	0.4	0.876	money-sup	1	0.167	0.286	0.918
coffee	0.6	0.6	0.6	0.996	nat-gas	0	0	0	0.705
copper	1	1	1	1	pet-chem	0	0	0	0.735
corn	0	0	0	0.972	potato	0	0	0	0.017
cotton	0	0	0	0.967	reserves	1	1	1	1
cpi	0.5	0.5	0.5	0.608	rice	0	0	0	0.997
crude	0.778	0.56	0.651	0.947	ship	0.556	0.294	0.385	0.904
dlr	0	0	0	0.984	soybean	0	0	0	0.942
earn	0.735	0.964	0.834	0.88	soy-meal	0	0	0	0.936
fuel	0	0	0	0.878	strat-metal	0	0	0	0.574
gas	0	0	0	0.811	sugar	1	0.2	0.333	0.976
gnp	0.667	0.4	0.5	0.99	tea	0	0	0	0.983
gold	0	0	0	0.973	trade	0.75	0.6	0.667	0.984
grain	0.5	0.25	0.333	0.968	veg-oil	0.5	0.333	0.4	0.959
groundnut	0	0	0	0.5	wheat	0.625	1	0.769	0.997
heat	0.5	1	0.667	0.999	wpi	0	0	0	0.975
instal-debt	0	0	0	0.5	yen	0	0	0	0.996
interest	0.667	0.571	0.615	0.811	zinc	0	0	0	0.568
ipi	0	0	0	0.945					
iron-steel	1	0.25	0.4	0.507					

Tabela A.3.7: Performance por classe com representação de documentos por Pentagramas

Apêndice B

K-NN: Resultados por classe com colecção R11

B.1 Performance obtida com o classificador K-NN utilizando a técnica do Terceiro Momento

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.498	0.944	0.652	0.929	livestock	0	0	0	0.697
alum	1	0.5	0.667	1	money-fx	0.5	0.25	0.333	0.874
bop	1	0.333	0.5	0.808	money-sup	0	0	0	0.873
carcass	0	0	0	0.834	nat-gas	0	0	0	0.853
cocoa	0	0	0	0.627	nzdlr	0	0	0	0.899
coffee	0	0	0	0.874	orange	0	0	0	0.638
copper	0	0	0	0.673	pet-chem	0	0	0	0.441
corn	0	0	0	0.963	platinum	0	0	0	0.614
cotton	0	0	0	0.575	potato	0	0	0	0.777
cpi	1	0.2	0.333	0.873	reserves	1	0.25	0.4	0.518
crude	0.529	0.391	0.45	0.884	rice	0	0	0	0.659
dlr	0	0	0	0.766	ship	0.375	0.214	0.273	0.919
earn	0.913	0.92	0.916	0.966	soybean	0	0	0	0.911
fuel	0	0	0	0.651	soy-meal	0	0	0	0.892
gas	0	0	0	1	soy-oil	0	0	0	0.956
gnp	0	0	0	0.876	strat-metal	0	0	0	0.443
gold	0.5	1	0.667	1	sugar	0	0	0	0.437
grain	0	0	0	0.878	tin	0	0	0	0.824
heat	0	0	0	0.902	trade	0.375	0.176	0.24	0.955
interest	0.355	0.44	0.393	0.923	veg-oil	0	0	0	0.844
ipi	0	0	0	0.688	wheat	0.667	0.5	0.571	0.886
iron-steel	0	0	0	0.523	wpi	1	0.5	0.667	1
jobs	0	0	0	0.751	yen	0	0	0	0.89
l-cattle	0	0	0	0.561	zinc	0	0	0	0.674
lei	0	0	0	0.793					

Tabela B.1.1: Performance por classe com representação de documentos por Palavras

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.413	0.772	0.538	0.876	lumber	0	0	0	0.891
alum	0	0	0	0.591	money-fx	0.636	0.226	0.333	0.774
barley	0	0	0	0.447	money-sup	0	0	0	0.857
bop	0.5	0.25	0.333	0.929	naphtha	0	0	0	0.818
carcass	0	0	0	0.771	nat-gas	0	0	0	0.737
cocoa	0	0	0	1	nkr	0	0	0	0.954
coffee	0	0	0	0.616	nzdlr	0	0	0	0.941
copper	0	0	0	0.504	orange	0	0	0	0.715
corn	0	0	0	0.906	pet-chem	0	0	0	0.369
cotton	0	0	0	0.805	potato	0	0	0	1
cpi	0	0	0	0.735	propane	0	0	0	0.785
crude	0.667	0.286	0.4	0.9	rapeseed	0	0	0	0.766
dlr	0	0	0	0.86	reserves	0	0	0	0.851
earn	0.716	0.892	0.794	0.935	rice	0	0	0	0.761
fuel	0	0	0	0.75	ship	0.444	0.308	0.364	0.898
gas	0	0	0	0.784	soybean	0	0	0	0.793
gnp	0	0	0	0.887	soy-oil	0	0	0	0.904
gold	0	0	0	0.448	strat-metal	1	0.333	0.5	0.78
grain	0	0	0	0.588	sugar	0	0	0	0.928
housing	0	0	0	0.145	trade	0.5	0.263	0.345	0.822
interest	0.667	0.526	0.588	0.951	veg-oil	0	0	0	0.726
ipi	0	0	0	0.74	wheat	0	0	0	0.806
iron-steel	0	0	0	0.735	wpi	0	0	0	0.844
jobs	0	0	0	0.848	yen	0	0	0	0.993
livestock	0	0	0	0.375					

Tabela B.1.2: Performance por classe com representação de documentos por Multi-palavras

B.1 Performance obtida com o classificador K-NN utilizando a técnica do Terceiro Momento

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.503	0.764	0.606	0.903	lumber	0	0	0	0.889
alum	1	0.143	0.25	0.714	money-fx	0.429	0.194	0.267	0.832
barley	0	0	0	0.448	money-sup	1	0.286	0.444	0.723
bop	0.5	0.25	0.333	0.971	naphtha	0	0	0	0.474
carcass	0	0	0	0.979	nat-gas	0	0	0	0.661
cocoa	1	1	1	1	nkr	0	0	0	0.739
coffee	0	0	0	0.583	nzdlr	0	0	0	0.645
copper	0	0	0	0.788	orange	0	0	0	0.995
corn	0	0	0	0.934	pet-chem	0	0	0	1
cotton	1	0.667	0.8	0.999	potato	1	1	1	1
cpi	0.5	0.25	0.333	0.555	propane	0	0	0	0.794
crude	0.75	0.321	0.45	0.864	rapeseed	0	0	0	0.54
dlr	0	0	0	0.697	reserves	0	0	0	0.848
earn	0.671	0.953	0.788	0.962	rice	1	0.333	0.5	0.894
fuel	0	0	0	0.486	ship	0.5	0.231	0.316	0.882
gas	0	0	0	0.962	soybean	0.5	0.5	0.5	0.953
gnp	0	0	0	0.84	soy-oil	0	0	0	0.869
gold	0.5	0.167	0.25	0.777	strat-metal	1	0.333	0.5	0.837
grain	0	0	0	0.512	sugar	0	0	0	0.803
housing	0	0	0	0.586	trade	0.333	0.158	0.214	0.841
interest	0.429	0.316	0.364	0.905	veg-oil	0	0	0	0.637
ipi	1	0.5	0.667	0.996	wheat	0.667	0.25	0.364	0.914
iron-steel	0	0	0	0.783	wpi	1	0.4	0.571	0.941
jobs	0	0	0	0.983	yen	0	0	0	0.995
livestock	0	0	0	0.348					

Tabela B.1.3: Performance por classe com representação de documentos por sequências dos primeiros 4 caracteres

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.622	0.617	0.62	0.869	jet	0	0	0	0.221
alum	1	0.75	0.857	1	jobs	0	0	0	0.746
barley	0	0	0	0.987	lead	0	0	0	0.61
bop	0.5	0.25	0.333	0.829	livestock	0	0	0	0.503
carcass	0	0	0	0.849	money-fx	0.583	0.35	0.438	0.825
cocoa	0	0	0	0.901	money-sup	0	0	0	0.504
coffee	0	0	0	0.903	nat-gas	0	0	0	0.837
copper	0	0	0	0.968	pet-chem	0	0	0	0.953
corn	0	0	0	0.925	potato	0	0	0	0.712
cotton	0.5	1	0.667	1	reserves	0	0	0	0.9
cpi	0	0	0	0.868	rice	1	1	1	1
crude	0.412	0.28	0.333	0.794	ship	0.333	0.118	0.174	0.761
dlr	0.5	0.5	0.5	0.678	soybean	0	0	0	0.768
earn	0.623	0.946	0.751	0.911	soy-meal	0	0	0	0.719
fuel	0	0	0	0.749	strat-metal	0	0	0	0.826
gas	0	0	0	0.397	sugar	0	0	0	0.749
gnp	0	0	0	0.924	tea	0	0	0	0.731
gold	0	0	0	0.212	trade	0.667	0.267	0.381	0.836
grain	0	0	0	0.88	veg-oil	0	0	0	0.673
groundnut	0	0	0	0.571	wheat	0.667	0.8	0.727	0.961
heat	0	0	0	0.989	wpi	0	0	0	0.378
instal-debt	0	0	0	0.564	yen	0	0	0	0.982
interest	0.4	0.19	0.258	0.825	zinc	0	0	0	0.868
ipi	0	0	0	0.988					
iron-steel	0	0	0	0.793					

Tabela B.1.4: Performance por classe com representação de documentos por sequências dos primeiros 5 caracteres

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.569	0.514	0.54	0.857	jobs	0	0	0	0.528
alum	0.75	0.75	0.75	0.999	lead	0	0	0	0.612
barley	1	1	1	1	lei	0	0	0	0.722
bop	0.25	0.2	0.222	0.853	livestock	0	0	0	0.609
carcass	0	0	0	0.985	meal-feed	0	0	0	0.793
cocoa	0.167	1	0.286	1	money-fx	0.5	0.37	0.426	0.849
coffee	1	0.2	0.333	0.76	money-sup	0	0	0	0.84
copper	0	0	0	0.996	nat-gas	0	0	0	0.488
corn	0	0	0	0.484	nkr	0	0	0	0.356
cotton	0	0	0	0.857	orange	0.5	1	0.667	0.998
cpi	0	0	0	0.548	palm-oil	0	0	0	0.998
crude	0.6	0.31	0.409	0.782	pet-chem	0	0	0	0.77
dlr	0	0	0	0.624	rapeseed	1	1	1	1
dmk	0	0	0	0.512	reserves	0	0	0	0.762
earn	0.573	0.946	0.713	0.911	rice	0	0	0	0.989
fuel	1	1	1	1	ship	0.5	0.111	0.182	0.761
gnp	0	0	0	0.88	soybean	0	0	0	1
gold	1	0.143	0.25	0.645	sugar	0.5	0.333	0.4	0.808
grain	0	0	0	0.636	tin	0	0	0	0.777
groundnut	0	0	0	0.54	trade	0.615	0.571	0.593	0.789
heat	0	0	0	0.858	veg-oil	0	0	0	0.911
income	0	0	0	0.948	wheat	0.286	0.286	0.286	0.981
interest	0.111	0.043	0.063	0.742	wpi	0	0	0	0.713
iron-steel	0	0	0	0.943	zinc	0	0	0	0.681
jet	0	0	0	0.242					

Tabela B.1.5: Performance por classe com representação de documentos por sequências dos primeiros 6 caracteres

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.452	0.537	0.491	0.797	lumber	0	0	0	0.91
alum	0	0	0	0.509	money-fx	0.429	0.194	0.267	0.822
barley	0	0	0	1	money-sup	0	0	0	0.676
bop	0	0	0	0.832	naphtha	0	0	0	0.42
carcass	0	0	0	0.963	nat-gas	0	0	0	0.577
cocoa	0.25	1	0.4	1	nkr	0	0	0	0.585
coffee	0.167	0.167	0.167	0.641	nzdlr	0	0	0	0.622
copper	1	0.25	0.4	0.895	orange	0	0	0	0.687
corn	1	0.2	0.333	0.833	pet-chem	0	0	0	0.563
cotton	1	0.667	0.8	0.921	potato	0	0	0	0.998
cpi	0	0	0	0.298	propane	0	0	0	0.703
crude	0.522	0.429	0.471	0.822	rapeseed	0	0	0	0.998
dlr	0	0	0	0.713	reserves	1	0.25	0.4	0.899
earn	0.622	0.901	0.736	0.89	rice	0	0	0	0.724
fuel	0	0	0	0.675	ship	0.286	0.154	0.2	0.844
gas	0.5	0.667	0.571	0.919	soybean	0	0	0	0.587
gnp	0	0	0	0.774	soy-oil	0	0	0	0.841
gold	1	0.167	0.286	0.575	strat-metal	1	0.333	0.5	0.87
grain	0	0	0	0.506	sugar	0	0	0	0.413
housing	0	0	0	0.278	trade	0.5	0.316	0.387	0.85
interest	0.333	0.316	0.324	0.904	veg-oil	0	0	0	0.236
ipi	0	0	0	0.971	wheat	0.333	0.125	0.182	0.773
iron-steel	0	0	0	0.549	wpi	0	0	0	0.721
jobs	1	1	1	1	yen	0	0	0	0.886
livestock	0	0	0	0.324					

Tabela B.1.6: Performance por classe com representação de documentos por sequências dos primeiros 4, 5 e 6 caracteres

B.1 Performance obtida com o classificador K-NN utilizando a técnica do Terceiro Momento

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.605	0.597	0.601	0.759	jet	0	0	0	0.822
alum	0	0	0	0.509	jobs	0.333	0.333	0.333	0.734
barley	1	0.5	0.667	0.995	lead	0.25	0.5	0.333	0.738
bop	0	0	0	0.708	livestock	0	0	0	0.14
carcass	0	0	0	0.923	money-fx	0.381	0.4	0.39	0.759
cocoa	0	0	0	0.673	money-sup	0	0	0	0.251
coffee	0.333	0.2	0.25	0.787	nat-gas	0	0	0	0.382
copper	0	0	0	0.194	pet-chem	0.333	0.333	0.333	0.848
corn	0	0	0	0.817	potato	0	0	0	0.924
cotton	0	0	0	0.609	reserves	0.333	0.333	0.333	0.901
cpi	0	0	0	0.757	rice	1	1	1	1
crude	0.476	0.4	0.435	0.736	ship	0.143	0.118	0.129	0.641
dlr	0	0	0	0.195	soybean	0	0	0	0.548
earn	0.66	0.779	0.715	0.768	soy-meal	0	0	0	0.371
fuel	0	0	0	0.614	strat-metal	0	0	0	0.555
gas	0	0	0	0.682	sugar	0	0	0	0.642
gnp	0	0	0	0.663	tea	0	0	0	0.698
gold	0	0	0	0.215	trade	0.211	0.267	0.235	0.828
grain	0	0	0	0.533	veg-oil	0	0	0	0.683
groundnut	0	0	0	0.604	wheat	0.333	0.4	0.364	0.747
heat	0.5	1	0.667	1	wpi	0	0	0	0.801
instal-debt	0	0	0	0.638	yen	0	0	0	0.956
interest	0.267	0.19	0.222	0.706	zinc	0	0	0	0.064
ipi	0	0	0	0.438					
iron-steel	0	0	0	0.441					

Tabela B.1.7: Performance por classe com representação de documentos por Pentagramas

B.2 Performance obtida com o classificador K-NN utilizando a técnica *Chi-Square*

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.781	0.863	0.82	0.973	livestock	0	0	0	0.533
alum	1	0.5	0.667	0.856	money-fx	0.682	0.625	0.652	0.908
bop	0.5	0.333	0.4	0.736	money-sup	0.8	0.8	0.8	0.816
carcass	0	0	0	0.932	nat-gas	1	0.5	0.667	0.99
cocoa	0	0	0	0.36	nzdlr	0	0	0	0.925
coffee	1	0.667	0.8	0.999	orange	1	0.667	0.8	0.739
copper	1	0.2	0.333	0.688	pet-chem	0	0	0	0.719
corn	0	0	0	0.993	platinum	0	0	0	0.726
cotton	0	0	0	0.425	potato	0	0	0	0.932
cpi	0.5	0.6	0.545	0.995	reserves	1	0.5	0.667	0.748
crude	0.765	0.565	0.65	0.942	rice	0	0	0	0.598
dlr	0	0	0	0.849	ship	0.417	0.714	0.526	0.925
earn	0.829	0.958	0.889	0.985	soybean	0	0	0	0.393
fuel	0	0	0	0.59	soy-meal	0	0	0	0.07
gas	1	1	1	1	soy-oil	0	0	0	0.643
gnp	0.5	0.333	0.4	0.811	strat-metal	0	0	0	0.598
gold	1	1	1	1	sugar	0	0	0	0.997
grain	0	0	0	0.826	tin	0	0	0	0.855
heat	0	0	0	0.997	trade	0.632	0.706	0.667	0.967
interest	0.536	0.6	0.566	0.934	veg-oil	0.667	0.667	0.667	0.975
ipi	0	0	0	0.63	wheat	0.833	0.625	0.714	0.921
iron-steel	0	0	0	0.884	wpi	0.333	0.5	0.4	0.999
jobs	0.5	0.5	0.5	0.983	yen	0	0	0	0.995
l-cattle	1	1	1	1	zinc	0	0	0	0.066
lei	0	0	0	0.257					

Tabela B.2.1: Performance por classe com representação de documentos por Palavras

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.498	0.946	0.653	0.91	lei	0	0	0	0.306
alum	0	0	0	0.854	livestock	0	0	0	0.464
barley	0	0	0	0.114	meal-feed	0	0	0	0.869
bop	0	0	0	0.995	money-fx	0.308	0.167	0.216	0.713
carcass	0	0	0	0.993	money-sup	1	0.125	0.222	0.85
cocoa	0.5	0.5	0.5	0.946	nat-gas	0.333	0.125	0.182	0.656
coffee	0	0	0	0.744	nzdlr	0	0	0	0.96
copper	0	0	0	0.744	orange	0	0	0	0.613
corn	0	0	0	0.814	pet-chem	0	0	0	0.668
cotton	0	0	0	0.281	platinum	0	0	0	0.72
cpi	0	0	0	0.644	potato	0	0	0	0.727
crude	0.769	0.37	0.5	0.95	propane	0	0	0	0.203
dlr	0	0	0	0.825	reserves	0.667	0.667	0.667	0.916
earn	0.905	0.883	0.894	0.959	ship	0.714	0.357	0.476	0.727
gas	1	0.5	0.667	0.857	soybean	0	0	0	0.865
gnp	0	0	0	0.912	soy-meal	0	0	0	0.49
gold	1	0.25	0.4	0.871	sugar	0	0	0	0.75
grain	1	0.333	0.5	0.945	tin	0	0	0	0.543
heat	0	0	0	0.766	trade	0.571	0.471	0.516	0.884
income	0	0	0	0.309	veg-oil	0	0	0	0.87
interest	0.615	0.421	0.5	0.908	wheat	0.8	0.571	0.667	0.997
ipi	0	0	0	0.639	wpi	0	0	0	0.364
iron-steel	0	0	0	0.537	yen	1	1	1	1
jobs	0	0	0	0.66					
lead	0	0	0	0.455					

Tabela B.2.2: Performance por classe com representação de documentos por Multi-palavras

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.724	0.854	0.784	0.958	lumber	0	0	0	0.931
alum	1	0.571	0.727	0.77	money-fx	0.778	0.452	0.571	0.851
barley	0	0	0	0.537	money-sup	0.667	0.286	0.4	0.871
bop	0	0	0	0.906	naphtha	0	0	0	0.056
carcass	0	0	0	0.956	nat-gas	0	0	0	0.559
cocoa	0	0	0	1	nkr	0	0	0	0.656
coffee	0	0	0	0.715	nzdlr	0	0	0	0.857
copper	1	0.25	0.4	0.987	orange	1	1	1	1
corn	0	0	0	0.762	pet-chem	0	0	0	0.117
cotton	0	0	0	0.625	potato	0	0	0	0.136
cpi	0.25	0.25	0.25	0.988	propane	0	0	0	0.717
crude	0.65	0.464	0.542	0.87	rapeseed	0	0	0	0.994
dlr	0.333	0.2	0.25	0.648	reserves	1	0.25	0.4	0.848
earn	0.706	0.953	0.811	0.959	rice	0	0	0	0.792
fuel	0	0	0	0.65	ship	0.643	0.692	0.667	0.961
gas	0	0	0	0.997	soybean	0	0	0	0.677
gnp	0.667	0.667	0.667	0.956	soy-oil	0	0	0	0.626
gold	0.8	0.667	0.727	0.997	strat-metal	1	0.333	0.5	0.805
grain	0	0	0	0.969	sugar	0.75	0.5	0.6	0.838
housing	0	0	0	0.096	trade	0.5	0.211	0.296	0.907
interest	0.455	0.526	0.488	0.964	veg-oil	0	0	0	0.626
ipi	0	0	0	0.996	wheat	0.667	0.5	0.571	0.943
iron-steel	0	0	0	0.877	wpi	1	0.2	0.333	0.919
jobs	0	0	0	0.995	yen	0	0	0	0.986
livestock	0	0	0	0.56					

Tabela B.2.3: Performance por classe com representação de documentos por sequências dos primeiros 4 caracteres

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.819	0.758	0.787	0.942	jet	0	0	0	0.777
alum	1	0.5	0.667	1	jobs	0	0	0	0.405
barley	0	0	0	0.989	lead	0	0	0	0.991
bop	0.5	0.25	0.333	0.848	livestock	0	0	0	0.446
carcass	0	0	0	0.974	money-fx	0.529	0.45	0.486	0.919
cocoa	0	0	0	0.908	money-sup	1	0.167	0.286	0.617
coffee	0.75	0.6	0.667	0.903	nat-gas	0	0	0	0.556
copper	0	0	0	1	pet-chem	0	0	0	0.697
corn	0.167	0.333	0.222	0.963	potato	0	0	0	0.822
cotton	0	0	0	0.487	reserves	0	0	0	0.801
cpi	0.25	0.5	0.333	0.968	rice	0	0	0	1
crude	0.818	0.36	0.5	0.917	ship	0.5	0.235	0.32	0.894
dlr	0	0	0	0.965	soybean	0	0	0	0.745
earn	0.657	0.982	0.787	0.93	soy-meal	0	0	0	0.161
fuel	0	0	0	0.711	strat-metal	0	0	0	0.415
gas	0	0	0	0.539	sugar	1	0.2	0.333	0.838
gnp	0	0	0	0.884	tea	0	0	0	0.606
gold	0	0	0	0.608	trade	1	0.333	0.5	0.835
grain	0	0	0	0.872	veg-oil	0	0	0	0.751
groundnut	0	0	0	0.772	wheat	0.556	1	0.714	0.999
heat	0	0	0	0.982	wpi	0	0	0	0.984
instal-debt	0	0	0	0.517	yen	0	0	0	0.979
interest	0.588	0.476	0.526	0.91	zinc	0	0	0	0.715
ipi	0	0	0	0.316					
iron-steel	1	0.25	0.4	0.729					

Tabela B.2.4: Performance por classe com representação de documentos por sequências dos primeiros 5 caracteres

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.776	0.674	0.721	0.923	jobs	0	0	0	0.679
alum	0.8	1	0.889	1	lead	0	0	0	0.629
barley	0	0	0	0.989	lei	0	0	0	0.552
bop	1	0.2	0.333	0.868	livestock	0	0	0	0.504
carcass	0	0	0	0.821	meal-feed	0	0	0	0.494
cocoa	0	0	0	0.995	money-fx	0.667	0.519	0.583	0.913
coffee	1	0.2	0.333	0.914	money-sup	0	0	0	0.911
copper	1	1	1	1	nat-gas	0	0	0	0.455
corn	0	0	0	0.739	nkr	0	0	0	0.998
cotton	0	0	0	0.683	orange	0	0	0	0.993
cpi	0.333	0.25	0.286	0.525	palm-oil	0	0	0	0.899
crude	0.789	0.517	0.625	0.869	pet-chem	0	0	0	0.89
dlr	0	0	0	0.929	rapeseed	1	1	1	1
dmk	0	0	0	0.912	reserves	0	0	0	0.997
earn	0.643	0.971	0.773	0.949	rice	0	0	0	0.855
fuel	0	0	0	0.998	ship	1	0.278	0.435	0.839
gnp	0	0	0	0.934	soybean	0	0	0	0.759
gold	0.5	0.143	0.222	0.787	sugar	0	0	0	0.844
grain	0	0	0	0.619	tin	0	0	0	0.66
groundnut	0	0	0	0.788	trade	0.727	0.571	0.64	0.904
heat	0	0	0	0.996	veg-oil	1	0.25	0.4	0.883
income	0	0	0	0.325	wheat	0.5	0.429	0.462	0.972
interest	0.323	0.435	0.37	0.81	wpi	1	0.333	0.5	0.996
iron-steel	0	0	0	0.343	zinc	0	0	0	0.668
jet	0	0	0	0.853					

Tabela B.2.5: Performance por classe com representação de documentos por sequências dos primeiros 6 caracteres

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.778	0.854	0.814	0.962	lumber	0	0	0	1
alum	1	0.714	0.833	0.787	money-fx	0.75	0.484	0.588	0.875
barley	0	0	0	0.558	money-sup	1	0.143	0.25	0.88
bop	0.333	0.25	0.286	0.98	naphtha	0	0	0	0.1
carcass	0	0	0	0.895	nat-gas	0	0	0	0.577
cocoa	1	1	1	1	nkr	0	0	0	0.695
coffee	1	0.5	0.667	0.768	nzdlr	0	0	0	0.925
copper	0	0	0	1	orange	1	1	1	1
corn	0	0	0	0.768	pet-chem	0	0	0	0.151
cotton	0	0	0	0.65	potato	0	0	0	0.145
cpi	0.6	0.75	0.667	0.995	propane	0	0	0	0.815
crude	0.813	0.464	0.591	0.897	rapeseed	1	0.5	0.667	0.997
dlr	0.25	0.2	0.222	0.797	reserves	0.5	0.25	0.333	0.899
earn	0.732	0.953	0.828	0.97	rice	0	0	0	0.833
fuel	0	0	0	0.997	ship	0.625	0.769	0.69	0.936
gas	0.333	0.333	0.333	0.958	soybean	0	0	0	0.692
gnp	0.5	0.333	0.4	0.922	soy-oil	0	0	0	0.783
gold	0.75	0.5	0.6	0.85	strat-metal	1	0.333	0.5	0.795
grain	0	0	0	0.974	sugar	0.75	0.5	0.6	0.821
housing	0	0	0	0.013	trade	0.636	0.368	0.467	0.92
interest	0.333	0.526	0.408	0.958	veg-oil	0	0	0	0.806
ipi	0.167	0.5	0.25	0.991	wheat	0.714	0.625	0.667	0.985
iron-steel	0.667	1	0.8	1	wpi	1	0.2	0.333	0.938
jobs	0	0	0	0.948	yen	0	0	0	0.997
livestock	0	0	0	0.823					

Tabela B.2.6: Performance por classe com representação de documentos por sequências dos primeiros 4, 5 e 6 caracteres

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.848	0.711	0.774	0.941	jet	0	0	0	0.78
alum	1	0.75	0.857	1	jobs	0	0	0	0.366
barley	0	0	0	0.178	lead	0	0	0	0.996
bop	0.333	0.25	0.286	0.89	livestock	0	0	0	0.269
carcass	0	0	0	0.998	money-fx	0.5	0.4	0.444	0.934
cocoa	0	0	0	0.906	money-sup	1	0.167	0.286	0.61
coffee	1	0.6	0.75	0.886	nat-gas	0	0	0	0.422
copper	0	0	0	0.439	pet-chem	0	0	0	0.707
corn	0	0	0	0.951	potato	0	0	0	0.9
cotton	0	0	0	0.859	reserves	0	0	0	0.814
cpi	0.5	0.5	0.5	0.973	rice	0	0	0	0.995
crude	0.643	0.36	0.462	0.926	ship	0.714	0.294	0.417	0.917
dlr	0.5	0.5	0.5	0.988	soybean	0	0	0	0.695
earn	0.65	0.968	0.778	0.942	soy-meal	0	0	0	0.166
fuel	0	0	0	0.673	strat-metal	0	0	0	0.842
gas	0	0	0	0.552	sugar	0.667	0.4	0.5	0.844
gnp	0	0	0	0.856	tea	0	0	0	0.638
gold	0	0	0	0.991	trade	1	0.333	0.5	0.816
grain	0	0	0	0.958	veg-oil	0	0	0	0.75
groundnut	0	0	0	0.734	wheat	0.444	0.8	0.571	0.995
heat	0	0	0	1	wpi	0	0	0	0.995
instal-debt	0	0	0	0.411	yen	0	0	0	0.975
interest	0.5	0.429	0.462	0.903	zinc	0	0	0	0.993
ipi	0	0	0	0.568					
iron-steel	1	0.25	0.4	0.742					

Tabela B.2.7: Performance por classe com representação de documentos por Pentagramas

B.3 Performance obtida com o classificador K-NN utilizando a técnica *Information Gain*

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.812	0.871	0.84	0.977	livestock	0	0	0	0.524
alum	0.5	0.5	0.5	0.849	money-fx	0.636	0.583	0.609	0.906
bop	1	0.333	0.5	0.734	money-sup	1	0.6	0.75	0.816
carcass	0	0	0	0.939	nat-gas	0	0	0	0.969
cocoa	0	0	0	0.361	nzdlr	0	0	0	0.889
coffee	1	1	1	1	orange	1	0.333	0.5	0.737
copper	1	0.4	0.571	0.844	pet-chem	0	0	0	0.717
corn	1	1	1	0.993	platinum	0	0	0	0.839
cotton	0	0	0	0.807	potato	0	0	0	0.932
cpi	0.5	0.6	0.545	0.996	reserves	1	0.25	0.4	0.707
crude	0.765	0.565	0.65	0.946	rice	0	0	0	0.585
dlr	0	0	0	0.846	ship	0.455	0.714	0.556	0.928
earn	0.829	0.958	0.889	0.987	soybean	0	0	0	0.392
fuel	1	0.333	0.5	0.569	soy-meal	0	0	0	0.068
gas	1	1	1	1	soy-oil	0	0	0	0.652
gnp	0	0	0	0.809	strat-metal	0	0	0	0.599
gold	0.5	1	0.667	1	sugar	0	0	0	0.993
grain	0	0	0	0.827	tin	0	0	0	0.844
heat	0	0	0	0.997	trade	0.682	0.882	0.769	0.971
interest	0.593	0.64	0.615	0.94	veg-oil	1	0.667	0.8	0.969
ipi	0	0	0	0.63	wheat	0.833	0.625	0.714	0.868
iron-steel	0	0	0	0.559	wpi	0.333	0.5	0.4	1
jobs	1	0.5	0.667	0.988	yen	0	0	0	0.995
l-cattle	1	1	1	1	zinc	0	0	0	0.066
lei	0	0	0	0.256					

Tabela B.3.1: Performance por classe com representação de documentos por Palavras

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.588	0.926	0.719	0.932	lei	0	0	0	0.99
alum	0	0	0	0.835	livestock	0	0	0	0.488
barley	0	0	0	0.094	meal-feed	0	0	0	0.893
bop	0	0	0	0.98	money-fx	0.081	0.125	0.098	0.728
carcass	0	0	0	0.998	money-sup	1	0.25	0.4	0.794
cocoa	0.333	0.5	0.4	0.947	nat-gas	0	0	0	0.627
coffee	0	0	0	0.471	nzdlr	0	0	0	0.899
copper	0	0	0	0.714	orange	0	0	0	0.709
corn	0	0	0	0.806	pet-chem	0	0	0	0.715
cotton	0	0	0	0.264	platinum	0	0	0	0.767
cpi	0	0	0	0.712	potato	0	0	0	0.855
crude	0.706	0.444	0.545	0.909	propane	0	0	0	0.278
dlr	0	0	0	0.785	reserves	0	0	0	0.877
earn	0.915	0.898	0.906	0.963	ship	0.364	0.286	0.32	0.746
gas	0	0	0	0.865	soybean	0	0	0	0.846
gnp	0	0	0	0.918	soy-meal	0	0	0	0.304
gold	0.5	0.25	0.333	0.839	sugar	0	0	0	0.548
grain	0	0	0	0.761	tin	0	0	0	0.466
heat	0	0	0	0.808	trade	0.6	0.353	0.444	0.889
income	0	0	0	0.173	veg-oil	0	0	0	0.727
interest	0.421	0.421	0.421	0.885	wheat	0.667	0.286	0.4	0.859
ipi	0	0	0	0.581	wpi	0	0	0	0.298
iron-steel	0	0	0	0.231	yen	1	1	1	1
jobs	0.2	0.333	0.25	0.793					
lead	0	0	0	0.322					

Tabela B.3.2: Performance por classe com representação de documentos por Multi-palavras

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.724	0.854	0.784	0.954	lumber	0	0	0	0.931
alum	1	0.571	0.727	0.769	money-fx	0.737	0.452	0.56	0.85
barley	0	0	0	0.536	money-sup	0.667	0.286	0.4	0.871
bop	0	0	0	0.909	naphtha	0	0	0	0.056
carcass	0	0	0	0.956	nat-gas	0	0	0	0.559
cocoa	0	0	0	1	nkr	0	0	0	0.654
coffee	0	0	0	0.715	nzdlr	0	0	0	0.86
copper	1	0.25	0.4	0.987	orange	1	1	1	1
corn	0	0	0	0.761	pet-chem	0	0	0	0.117
cotton	0	0	0	0.623	potato	0	0	0	0.133
cpi	0.4	0.5	0.444	0.989	propane	0	0	0	0.71
crude	0.7	0.5	0.583	0.869	rapeseed	0	0	0	0.994
dlr	0	0	0	0.647	reserves	1	0.25	0.4	0.845
earn	0.709	0.953	0.813	0.959	rice	0	0	0	0.791
fuel	0	0	0	0.652	ship	0.714	0.769	0.741	0.962
gas	0	0	0	0.998	soybean	0	0	0	0.689
gnp	0.667	0.667	0.667	0.955	soy-oil	0	0	0	0.626
gold	0.8	0.667	0.727	0.997	strat-metal	1	0.333	0.5	0.805
grain	0	0	0	0.969	sugar	0.75	0.5	0.6	0.837
housing	0	0	0	0.094	trade	0.6	0.316	0.414	0.908
interest	0.455	0.526	0.488	0.964	veg-oil	0	0	0	0.624
ipi	0	0	0	0.996	wheat	0.667	0.5	0.571	0.942
iron-steel	0	0	0	0.876	wpi	1	0.2	0.333	0.919
jobs	0	0	0	0.995	yen	0	0	0	0.986
livestock	0	0	0	0.56					

Tabela B.3.3: Performance por classe com representação de documentos por sequências dos primeiros 4 caracteres

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.819	0.758	0.787	0.942	jet	0	0	0	0.777
alum	1	0.5	0.667	1	jobs	0	0	0	0.405
barley	0	0	0	0.989	lead	0	0	0	0.991
bop	0.5	0.25	0.333	0.848	livestock	0	0	0	0.446
carcass	0	0	0	0.974	money-fx	0.529	0.45	0.486	0.919
cocoa	0	0	0	0.908	money-sup	1	0.167	0.286	0.617
coffee	0.75	0.6	0.667	0.903	nat-gas	0	0	0	0.556
copper	0	0	0	1	pet-chem	0	0	0	0.697
corn	0.167	0.333	0.222	0.963	potato	0	0	0	0.822
cotton	0	0	0	0.487	reserves	0	0	0	0.801
cpi	0.25	0.5	0.333	0.968	rice	0	0	0	1
crude	0.818	0.36	0.5	0.917	ship	0.5	0.235	0.32	0.894
dlr	0	0	0	0.965	soybean	0	0	0	0.745
earn	0.657	0.982	0.787	0.93	soy-meal	0	0	0	0.161
fuel	0	0	0	0.711	strat-metal	0	0	0	0.415
gas	0	0	0	0.539	sugar	1	0.2	0.333	0.838
gnp	0	0	0	0.884	tea	0	0	0	0.606
gold	0	0	0	0.608	trade	1	0.333	0.5	0.835
grain	0	0	0	0.872	veg-oil	0	0	0	0.751
groundnut	0	0	0	0.772	wheat	0.556	1	0.714	0.999
heat	0	0	0	0.982	wpi	0	0	0	0.984
instal-debt	0	0	0	0.517	yen	0	0	0	0.979
interest	0.588	0.476	0.526	0.91	zinc	0	0	0	0.715
ipi	0	0	0	0.316					
iron-steel	1	0.25	0.4	0.729					

Tabela B.3.4: Performance por classe com representação de documentos por sequências dos primeiros 5 caracteres

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.776	0.674	0.721	0.923	jobs	0	0	0	0.679
alum	0.8	1	0.889	1	lead	0	0	0	0.629
barley	0	0	0	0.989	lei	0	0	0	0.552
bop	1	0.2	0.333	0.868	livestock	0	0	0	0.504
carcass	0	0	0	0.821	meal-feed	0	0	0	0.494
cocoa	0	0	0	0.995	money-fx	0.667	0.519	0.583	0.913
coffee	1	0.2	0.333	0.914	money-sup	0	0	0	0.911
copper	1	1	1	1	nat-gas	0	0	0	0.455
corn	0	0	0	0.739	nkr	0	0	0	0.998
cotton	0	0	0	0.681	orange	0	0	0	0.993
cpi	0.333	0.25	0.286	0.525	palm-oil	0	0	0	0.899
crude	0.789	0.517	0.625	0.869	pet-chem	0	0	0	0.89
dlr	0	0	0	0.929	rapeseed	1	1	1	1
dmk	0	0	0	0.912	reserves	0	0	0	0.997
earn	0.643	0.971	0.773	0.949	rice	0	0	0	0.855
fuel	0	0	0	0.998	ship	1	0.278	0.435	0.839
gnp	0	0	0	0.934	soybean	0	0	0	0.759
gold	0.5	0.143	0.222	0.79	sugar	0	0	0	0.843
grain	0	0	0	0.619	tin	0	0	0	0.66
groundnut	0	0	0	0.788	trade	0.727	0.571	0.64	0.904
heat	0	0	0	0.995	veg-oil	1	0.25	0.4	0.883
income	0	0	0	0.325	wheat	0.5	0.429	0.462	0.972
interest	0.323	0.435	0.37	0.811	wpi	1	0.333	0.5	0.996
iron-steel	0	0	0	0.343	zinc	0	0	0	0.668
jet	0	0	0	0.853					

Tabela B.3.5: Performance por classe com representação de documentos por sequências dos primeiros 6 caracteres

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.778	0.854	0.814	0.962	lumber	0	0	0	1
alum	1	0.714	0.833	0.787	money-fx	0.75	0.484	0.588	0.875
barley	0	0	0	0.558	money-sup	1	0.143	0.25	0.88
bop	0.333	0.25	0.286	0.98	naphtha	0	0	0	0.1
carcass	0	0	0	0.895	nat-gas	0	0	0	0.577
cocoa	1	1	1	1	nkr	0	0	0	0.695
coffee	1	0.5	0.667	0.768	nzdlr	0	0	0	0.925
copper	0	0	0	1	orange	1	1	1	1
corn	0	0	0	0.768	pet-chem	0	0	0	0.151
cotton	0	0	0	0.65	potato	0	0	0	0.145
cpi	0.6	0.75	0.667	0.995	propane	0	0	0	0.815
crude	0.813	0.464	0.591	0.897	rapeseed	1	0.5	0.667	0.997
dlr	0.25	0.2	0.222	0.797	reserves	0.5	0.25	0.333	0.899
earn	0.732	0.953	0.828	0.97	rice	0	0	0	0.833
fuel	0	0	0	0.997	ship	0.625	0.769	0.69	0.936
gas	0.333	0.333	0.333	0.958	soybean	0	0	0	0.692
gnp	0.5	0.333	0.4	0.922	soy-oil	0	0	0	0.783
gold	0.75	0.5	0.6	0.85	strat-metal	1	0.333	0.5	0.795
grain	0	0	0	0.974	sugar	0.75	0.5	0.6	0.821
housing	0	0	0	0.013	trade	0.636	0.368	0.467	0.92
interest	0.333	0.526	0.408	0.958	veg-oil	0	0	0	0.806
ipi	0.167	0.5	0.25	0.991	wheat	0.714	0.625	0.667	0.985
iron-steel	0.667	1	0.8	1	wpi	1	0.2	0.333	0.938
jobs	0	0	0	0.948	yen	0	0	0	0.997
livestock	0	0	0	0.823					

Tabela B.3.6: Performance por classe com representação de documentos por sequências dos primeiros 4, 5 e 6 caracteres

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.848	0.711	0.774	0.941	jet	0	0	0	0.78
alum	1	0.75	0.857	1	jobs	0	0	0	0.366
barley	0	0	0	0.178	lead	0	0	0	0.996
bop	0.333	0.25	0.286	0.89	livestock	0	0	0	0.269
carcass	0	0	0	0.998	money-fx	0.5	0.4	0.444	0.934
cocoa	0	0	0	0.906	money-sup	1	0.167	0.286	0.61
coffee	1	0.6	0.75	0.886	nat-gas	0	0	0	0.422
copper	0	0	0	0.439	pet-chem	0	0	0	0.707
corn	0	0	0	0.951	potato	0	0	0	0.9
cotton	0	0	0	0.859	reserves	0	0	0	0.814
cpi	0.5	0.5	0.5	0.973	rice	0	0	0	0.995
crude	0.643	0.36	0.462	0.926	ship	0.714	0.294	0.417	0.917
dlr	0.5	0.5	0.5	0.988	soybean	0	0	0	0.695
earn	0.65	0.968	0.778	0.942	soy-meal	0	0	0	0.166
fuel	0	0	0	0.673	strat-metal	0	0	0	0.842
gas	0	0	0	0.552	sugar	0.667	0.4	0.5	0.844
gnp	0	0	0	0.856	tea	0	0	0	0.638
gold	0	0	0	0.991	trade	1	0.333	0.5	0.816
grain	0	0	0	0.958	veg-oil	0	0	0	0.75
groundnut	0	0	0	0.734	wheat	0.444	0.8	0.571	0.995
heat	0	0	0	1	wpi	0	0	0	0.995
instal-debt	0	0	0	0.411	yen	0	0	0	0.975
interest	0.5	0.429	0.462	0.903	zinc	0	0	0	0.993
ipi	0	0	0	0.568					
iron-steel	1	0.25	0.4	0.742					

Tabela B.3.7: Performance por classe com representação de documentos por Pentagramas

Apêndice C

RIPPER: Resultados por classe com a colecção R11

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.866	0.887	0.876	0.938	livestock	0.667	0.667	0.667	0.758
alum	1	1	1	1	money-fx	0.55	0.458	0.5	0.78
bop	0.667	0.667	0.667	0.829	money-sup	0.4	0.4	0.4	0.76
carcass	0	0	0	0.5	nat-gas	0.333	0.5	0.4	0.748
cocoa	1	1	1	1	nzdlr	0	0	0	0.278
coffee	1	1	1	1	orange	1	1	1	1
copper	1	1	1	1	pet-chem	0	0	0	0.712
corn	1	1	1	1	platinum	0	0	0	0.277
cotton	1	1	1	1	potato	0	0	0	0.5
cpi	0.75	0.6	0.667	0.907	reserves	1	0.25	0.4	0.625
crude	0.762	0.696	0.727	0.957	rice	0.667	1	0.8	0.999
dlr	0.8	0.571	0.667	0.861	ship	0.421	0.571	0.485	0.788
earn	0.859	0.92	0.888	0.917	soybean	0.5	0.667	0.571	0.832
fuel	0	0	0	0.653	soy-meal	0	0	0	0.997
gas	0	0	0	0.498	soy-oil	0	0	0	0.278
gnp	0.5	0.667	0.571	0.921	strat-metal	0	0	0	0.261
gold	0.167	1	0.286	0.996	sugar	1	1	1	1
grain	0.4	0.667	0.5	0.751	tin	1	1	1	1
heat	0	0	0	0.499	trade	1	0.706	0.828	0.871
interest	0.778	0.56	0.651	0.818	veg-oil	1	1	1	1
ipi	0.667	0.4	0.5	0.766	wheat	1	0.75	0.857	0.875
iron-steel	0.5	0.75	0.6	0.812	wpi	0.5	0.5	0.5	0.748
jobs	1	1	1	1	yen	0.25	1	0.4	0.997
l-cattle	0	0	0	0.998	zinc	0.5	1	0.667	0.999
lei	0	0	0	0.779					

Tabela C.1: Performance por classe com representação de documentos por Palavras

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.866	0.887	0.876	0.938	livestock	0.667	0.667	0.667	0.758
alum	1	1	1	1	money-fx	0.55	0.458	0.5	0.78
bop	0.667	0.667	0.667	0.829	money-sup	0.4	0.4	0.4	0.76
carcass	0	0	0	0.5	nat-gas	0.333	0.5	0.4	0.748
cocoa	1	1	1	1	nzdlr	0	0	0	0.278
coffee	1	1	1	1	orange	1	1	1	1
copper	1	1	1	1	pet-chem	0	0	0	0.712
corn	1	1	1	1	platinum	0	0	0	0.277
cotton	1	1	1	1	potato	0	0	0	0.5
cpi	0.75	0.6	0.667	0.907	reserves	1	0.25	0.4	0.625
crude	0.762	0.696	0.727	0.957	rice	0.667	1	0.8	0.999
dlr	0.8	0.571	0.667	0.861	ship	0.421	0.571	0.485	0.788
earn	0.859	0.92	0.888	0.917	soybean	0.5	0.667	0.571	0.832
fuel	0	0	0	0.653	soy-meal	0	0	0	0.997
gas	0	0	0	0.498	soy-oil	0	0	0	0.278
gnp	0.5	0.667	0.571	0.921	strat-metal	0	0	0	0.261
gold	0.167	1	0.286	0.996	sugar	1	1	1	1
grain	0.4	0.667	0.5	0.751	tin	1	1	1	1
heat	0	0	0	0.499	trade	1	0.706	0.828	0.871
interest	0.778	0.56	0.651	0.818	veg-oil	1	1	1	1
ipi	0.667	0.4	0.5	0.766	wheat	1	0.75	0.857	0.875
iron-steel	0.5	0.75	0.6	0.812	wpi	0.5	0.5	0.5	0.748
jobs	1	1	1	1	yen	0.25	1	0.4	0.997
l-cattle	0	0	0	0.998	zinc	0.5	1	0.667	0.999
lei	0	0	0	0.779					

Tabela C.1.2: Performance por classe com representação de documentos por Multi-palavras

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.858	0.935	0.895	0.944	lumber	0	0	0	0.499
alum	1	0.571	0.727	0.786	money-fx	0.842	0.516	0.64	0.962
barley	1	1	1	1	money-sup	1	0.143	0.25	0.774
bop	0.5	0.25	0.333	0.461	naphtha	0	0	0	0.5
carcass	1	1	1	1	nat-gas	0	0	0	0.671
cocoa	1	1	1	1	nkr	0	0	0	0.5
coffee	0.857	1	0.923	0.999	nzdlr	0	0	0	0.288
copper	1	1	1	1	orange	0.5	1	0.667	0.999
corn	1	0.8	0.889	0.9	pet-chem	0	0	0	0.68
cotton	1	1	1	1	potato	0	0	0	0.789
cpi	0	0	0	0.787	propane	0	0	0	0.5
crude	0.864	0.679	0.76	0.95	rapeseed	1	1	1	1
dlr	0.4	0.4	0.4	0.77	reserves	0.75	0.75	0.75	0.874
earn	0.806	0.92	0.859	0.905	rice	0.5	1	0.667	0.997
fuel	0	0	0	0.485	ship	0.526	0.769	0.625	0.874
gas	0.5	1	0.667	0.998	soybean	0	0	0	0.853
gnp	0	0	0	0.613	soy-oil	0	0	0	0.485
gold	0.571	0.667	0.615	0.926	strat-metal	0	0	0	0.595
grain	0	0	0	0.734	sugar	0.667	0.667	0.667	0.831
housing	0	0	0	0.768	trade	0.647	0.579	0.611	0.845
interest	0.519	0.737	0.609	0.925	veg-oil	1	1	1	1
ipi	0	0	0	0.79	wheat	1	1	1	1
iron-steel	0	0	0	0.634	wpi	0	0	0	0.489
jobs	1	1	1	1	yen	0.5	1	0.667	0.999
livestock	1	0.5	0.667	0.64					

Tabela C.1.3: Performance por classe com representação de documentos por sequências dos primeiros 4 caracteres

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.83	0.886	0.857	0.924	jet	0	0	0	0.5
alum	0.667	1	0.8	0.998	jobs	1	0.667	0.8	0.833
barley	0	0	0	0.5	lead	0	0	0	0.724
bop	0.2	0.25	0.222	0.62	livestock	1	1	1	1
carcass	0	0	0	0.489	money-fx	0.647	0.55	0.595	0.856
cocoa	1	1	1	1	money-sup	0.667	0.333	0.444	0.847
coffee	1	1	1	1	nat-gas	1	0.333	0.5	0.659
copper	1	1	1	1	pet-chem	0	0	0	0.611
corn	0	0	0	0.833	potato	0	0	0	0.283
cotton	1	0.5	0.667	0.75	reserves	0.667	0.667	0.667	0.755
cpi	0.2	0.5	0.286	0.636	rice	0	0	0	0.721
crude	0.7	0.56	0.622	0.886	ship	0.556	0.588	0.571	0.784
dlr	0	0	0	0.635	soybean	0	0	0	0.436
earn	0.794	0.883	0.836	0.886	soy-meal	0	0	0	0.784
fuel	0.333	0.25	0.286	0.46	strat-metal	0	0	0	0.282
gas	0	0	0	0.489	sugar	1	0.6	0.75	0.899
gnp	0.667	0.4	0.5	0.667	tea	0	0	0	0.784
gold	0.5	0.5	0.5	0.891	trade	0.467	0.467	0.467	0.71
grain	0.333	0.5	0.4	0.937	veg-oil	1	0.333	0.5	0.664
groundnut	0	0	0	0.5	wheat	0.6	0.6	0.6	0.996
heat	1	1	1	1	wpi	1	1	1	1
instal-debt	0	0	0	0.5	yen	0	0	0	0.5
interest	0.737	0.667	0.7	0.82	zinc	0	0	0	0.999
ipi	0.5	1	0.667	0.999					
iron-steel	1	0.75	0.857	0.945					

Tabela C.1.4: Performance por classe com representação de documentos por sequências dos primeiros 5 caracteres

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.844	0.75	0.794	0.877	jobs	0.5	0.667	0.571	0.832
alum	1	0.75	0.857	0.814	lead	1	0.333	0.5	0.672
barley	1	1	1	1	lei	0	0	0	0.76
bop	0	0	0	0.703	livestock	1	0.5	0.667	0.873
carcass	1	0.333	0.5	0.667	meal-feed	0	0	0	0.5
cocoa	0	0	0	0.242	money-fx	0.633	0.704	0.667	0.89
coffee	0.8	0.8	0.8	0.95	money-sup	0	0	0	0.636
copper	0	0	0	0.5	nat-gas	1	1	1	1
corn	0	0	0	0.452	nkr	0	0	0	0.258
cotton	1	1	1	1	orange	1	1	1	1
cpi	1	0.25	0.4	0.622	palm-oil	0	0	0	0.258
crude	0.741	0.69	0.714	0.802	pet-chem	0	0	0	0.664
dlr	0.167	0.167	0.167	0.785	rapeseed	1	1	1	1
dmk	0	0	0	0.5	reserves	0	0	0	0.76
earn	0.681	0.912	0.78	0.856	rice	0	0	0	0.258
fuel	0	0	0	0.984	ship	0.9	0.5	0.643	0.878
gnp	0	0	0	0.381	soybean	0	0	0	0.248
gold	1	0.857	0.923	0.892	sugar	0.333	0.333	0.333	0.835
grain	0	0	0	0.403	tin	0	0	0	0.245
groundnut	0	0	0	0.258	trade	0.529	0.643	0.581	0.831
heat	0	0	0	0.5	veg-oil	0.6	0.75	0.667	0.813
income	0	0	0	0.76	wheat	0.5	0.429	0.462	0.713
interest	0.55	0.478	0.512	0.832	wpi	0	0	0	0.593
iron-steel	0	0	0	0.258	zinc	0	0	0	0.593
jet	0	0	0	0.5					

Tabela C.1.5: Performance por classe com representação de documentos por sequências dos primeiros 6 caracteres

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.819	0.919	0.866	0.937	lumber	0	0	0	0.499
alum	1	0.571	0.727	0.786	money-fx	0.652	0.484	0.556	0.872
barley	1	1	1	1	money-sup	1	0.286	0.444	0.713
bop	0	0	0	0.414	naphtha	0	0	0	0.5
carcass	0	0	0	0.5	nat-gas	1	0.25	0.4	0.624
cocoa	1	1	1	1	nkr	0	0	0	0.5
coffee	0.857	1	0.923	0.999	nzdlr	0	0	0	0.496
copper	1	1	1	1	orange	0.5	1	0.667	0.999
corn	1	0.8	0.889	0.858	pet-chem	0	0	0	0.686
cotton	1	1	1	1	potato	1	1	1	1
cpi	0.25	0.5	0.333	0.745	propane	0	0	0	0.5
crude	0.933	0.5	0.651	0.896	rapeseed	0	0	0	0.982
dlr	0	0	0	0.615	reserves	1	0.25	0.4	0.469
earn	0.844	0.92	0.88	0.916	rice	0.5	1	0.667	0.997
fuel	0	0	0	0.477	ship	0.48	0.923	0.632	0.933
gas	1	0.667	0.8	0.999	soybean	0.667	1	0.8	0.999
gnp	0.25	0.667	0.364	0.925	soy-oil	0	0	0	0.48
gold	0.714	0.833	0.769	0.915	strat-metal	0	0	0	0.651
grain	1	1	1	1	sugar	0.667	0.667	0.667	0.831
housing	0	0	0	0.49	trade	0.6	0.474	0.529	0.775
interest	0.52	0.684	0.591	0.899	veg-oil	1	1	1	1
ipi	0	0	0	0.799	wheat	1	1	1	1
iron-steel	0.25	0.5	0.333	0.646	wpi	0	0	0	0.779
jobs	1	1	1	1	yen	0	0	0	0.489
livestock	1	0.5	0.667	0.649					

Tabela C.1.6: Performance por classe com representação de documentos por sequências dos primeiros 4, 5 e 6 caracteres

Classe	Precisão	*Recall*	F- Measure	Área ROC	Classe	Precisão	*Recall*	F- Measure	Área ROC
acq	0.834	0.846	0.84	0.922	jet	0	0	0	0.5
alum	0.667	1	0.8	0.998	jobs	1	0.667	0.8	0.833
bop	0	0	0	0.652	lead	0	0	0	0.712
carcass	0	0	0	0.496	livestock	1	1	1	1
cocoa	1	1	1	1	money-fx	0.696	0.8	0.744	0.935
coffee	1	1	1	1	money-sup	1	0.333	0.5	0.842
copper	1	1	1	1	nat-gas	0	0	0	0.388
corn	1	0.333	0.5	0.737	pet-chem	0	0	0	0.601
cotton	1	0.5	0.667	0.75	potato	1	1	1	1
cpi	0.25	0.5	0.333	0.885	reserves	0	0	0	0.789
crude	0.696	0.64	0.667	0.882	rice	0	0	0	0.763
dlr	0	0	0	0.629	ship	0.75	0.529	0.621	0.844
earn	0.782	0.905	0.839	0.889	soybean	0	0	0	0.393
fuel	0.333	0.25	0.286	0.453	soy-meal	0	0	0	0.775
gas	0	0	0	0.486	strat-metal	0	0	0	0.273
gnp	0.571	0.8	0.667	0.852	sugar	1	0.8	0.889	0.9
gold	0.5	0.5	0.5	0.636	tea	0	0	0	0.5
grain	0.333	0.5	0.4	0.811	trade	0.588	0.667	0.625	0.942
groundnut	0	0	0	0.5	veg-oil	1	0.333	0.5	0.667
heat	1	1	1	1	wheat	0.6	0.6	0.6	0.996
instal-debt	0	0	0	0.5	wpi	0.5	1	0.667	0.998
interest	0.647	0.524	0.579	0.745	yen	0	0	0	0.5
ipi	1	1	1	1	zinc	0	0	0	0.998
iron-steel	1	0.75	0.857	0.941					

Tabela C.1.7: Performance por classe com representação de documentos por Pentagramas

Apêndice D

SVM: Resultados por classe com a colecção R12

D.1 Performance obtida utilizando a técnica Terceiro Momento

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.936	0.954	0.945	0.968	146	1	2	0	2	1	0	1	0	0
earn	0.970	0.970	0.970	0.979	3	194	1	0	1	1	0	0	0	0
money-fx	0.733	0.815	0.772	0.944	1	1	22	1	0	0	2	0	0	0
grain	0.667	0.857	0.750	0.995	0	0	0	6	0	0	0	0	1	0
crude	0.800	0.667	0.7257	0.953	3	2	0	0	24	0	1	4	2	0
trade	0.842	0.842	0.842	0.956	1	2	0	0	0	16	0	0	0	0
interest	0.833	0.714	0.769	0.984	0	0	5	0	0	0	15	0	0	0
ship	0.583	0.636	0.609	0.980	1	0	0	0	3	0	0	7	0	0
wheat	0.583	0.700	0.636	0.982	1	0	0	2	0	0	0	0	7	0
corn	1	0.600	0.75	1	0	0	0	0	0	0	0	0	2	3

Tabela D.1.1: Performance por classe e Matriz de confusão para Palavras

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.818	0.963	0.884	0.948	130	2	0	0	0	1	0	1	0	1
earn	0.961	0.948	0.954	0.971	8	199	0	0	1	0	0	1	1	0
money-fx	0.700	0.438	0.538	0.858	3	3	14	0	1	3	8	0	0	0
grain	0.500	0.167	0.250	0.767	2	0	0	1	0	0	0	0	2	1
crude	0.767	0.742	0.754	0.924	5	1	0	0	23	1	0	1	0	0
trade	0.529	0.600	0.563	0.962	3	0	0	0	1	9	0	0	0	2
interest	0.600	0.480	0.533	0.851	4	1	6	0	0	2	12	0	0	0
ship	0.813	0.684	0.743	0.967	1	0	0	1	4	0	0	13	0	0
wheat	0.600	0.667	0.632	0.880	2	0	0	0	0	0	0	0	6	1
corn	0	0	0.883	0.883	1	1	0	0	0	1	0	0	1	0

Tabela D.1.2: Performance por classe e Matriz de confusão para Multi-palavras

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.913	0.942	0.927	0.96	147	7	0	0	0	2	0	0	0	0
earn	0.949	0.971	0.96	0.968	3	204	0	0	2	0	1	0	0	0
money-fx	0.852	0.719	0.78	0.928	2	2	23	0	0	2	3	0	0	0
grain	0.667	1	0.8	0.999	0	0	0	2	0	0	0	0	0	0
crude	0.696	0.593	0.64	0.928	4	0	1	0	16	0	0	5	1	0
trade	0.647	0.733	0.688	0.939	2	0	1	0	1	11	0	0	0	0
interest	0.81	0.708	0.756	0.959	3	1	2	0	0	1	17	0	0	0
ship	0.615	0.727	0.667	0.99	0	0	0	0	3	0	0	8	0	0
wheat	0.833	0.714	0.769	0.998	0	1	0	1	0	0	0	0	5	0
corn	0.913	0.942	0.927	0.96	0	0	0	0	1	1	0	0	0	1

Tabela D.1.3: Performance por classe e Matriz de confusão para sequências de 4-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.923	0.87	0.896	0.939	120	17	0	0	1	0	0	0	0	0
earn	0.864	0.966	0.912	0.916	6	228	0	0	1	0	1	0	0	0
money-fx	0.929	0.565	0.703	0.959	0	7	13	0	0	1	2	0	0	0
grain	0.8	0.5	0.615	0.903	1	1	0	4	0	1	0	1	0	0
crude	0.75	0.75	0.75	0.939	2	1	0	0	18	1	0	2	0	0
trade	0.667	0.769	0.714	0.982	1	2	0	0	0	10	0	0	0	0
interest	0.842	0.667	0.744	0.89	0	6	1	0	0	1	16	0	0	0
ship	0.5	0.333	0.4	0.979	0	2	0	0	4	0	0	3	0	0
wheat	1	0.875	0.933	1	0	0	0	1	0	0	0	0	7	0
corn	1	0.5	0.667	0.994	0	0	0	0	0	1	0	0	0	1

Tabela D.1.4: Performance por classe e Matriz de confusão para sequências de 5-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.883	0.883	0.883	0.934	121	14	0	0	1	0	1	0	0	0
earn	0.851	0.981	0.912	0.927	3	206	0	0	0	0	1	0	0	0
money-fx	0.864	0.655	0.745	0.932	2	2	19	0	0	2	4	0	0	0
grain	0.333	0.1	0.154	0.856	0	5	0	1	0	1	0	1	1	1
crude	0.789	0.577	0.667	0.887	5	4	0	0	15	2	0	0	0	0
trade	0.739	0.81	0.773	0.948	1	2	0	0	0	17	1	0	0	0
interest	0.636	0.583	0.609	0.851	1	5	3	0	0	1	14	0	0	0
ship	0.8	0.333	0.471	0.78	3	2	0	0	3	0	0	4	0	0
wheat	0.5	0.4	0.444	0.987	0	1	0	1	0	0	0	0	2	1
corn	0.333	0.167	0.222	0.605	1	1	0	1	0	0	1	0	1	1

Tabela D.1.5: Performance por classe e Matriz de confusão para sequências de 6-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.839	0.909	0.872	0.924	130	8	1	0	2	0	1	1	0	0
earn	0.927	0.958	0.942	0.955	6	203	1	0	1	0	1	0	0	0
money-fx	0.5	0.48	0.49	0.88	2	3	12	0	1	1	6	0	0	0
grain	0.4	0.286	0.333	0.971	3	1	0	2	0	0	0	1	0	0
crude	0.68	0.63	0.654	0.877	4	1	0	1	17	1	0	2	1	0
trade	0.778	0.609	0.683	0.948	2	0	2	1	2	14	0	2	0	0
interest	0.652	0.484	0.556	0.885	5	1	8	0	0	2	15	0	0	0
ship	0.4	0.4	0.4	0.881	2	2	0	0	2	0	0	4	0	0
wheat	0.6	0.6	0.6	0.852	1	0	0	0	0	0	0	0	3	1
corn	0.5	0.333	0.4	0.987	0	0	0	1	0	0	0	0	1	1

Tabela D.1.6: Performance por classe e Matriz de confusão para sequências de 4, 5 e 6-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.898	0.827	0.861	0.915	115	21	1	1	0	1	0	0	0	0
earn	0.856	0.972	0.91	0.928	4	208	1	0	1	0	0	0	0	0
money-fx	0.7	0.519	0.596	0.853	1	3	14	1	0	0	7	0	1	0
grain	0.4	0.667	0.5	0.982	0	1	0	4	0	1	0	0	0	0
crude	0.821	0.657	0.73	0.902	3	2	0	0	23	1	1	4	1	0
trade	0.75	0.571	0.649	0.96	2	4	2	0	0	12	0	1	0	0
interest	0.556	0.714	0.625	0.918	0	1	2	0	0	1	10	0	0	0
ship	0.643	0.474	0.545	0.893	3	3	0	0	4	0	0	9	0	0
wheat	0.571	0.571	0.571	0.99	0	0	0	3	0	0	0	0	4	0
corn	0	0	0	0.502	0	0	0	1	0	0	0	0	1	0

Tabela D.1.7: Performance por classe e Matriz de confusão para Pentagramas

D.2 Performance obtida utilizando a técnica *Chi Square*

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.975	0.965	0.97	0.981	193	4	0	0	1	1	0	0	0	1
earn	0.905	0.987	0.944	0.972	1	152	0	0	1	0	0	0	0	0
money-fx	0.923	0.889	0.906	0.974	0	1	24	0	0	0	2	0	0	0
grain	0.8	0.571	0.667	0.984	0	2	0	4	0	0	0	0	1	0
crude	0.893	0.676	0.769	0.96	2	4	0	0	25	3	0	2	1	0
trade	0.789	0.833	0.811	0.942	1	2	0	0	0	15	0	0	0	0
interest	0.882	0.75	0.811	0.981	1	1	2	0	0	0	15	1	0	0
ship	0.75	0.818	0.783	0.925	0	1	0	0	1	0	0	9	0	0
wheat	0.667	0.8	0.727	0.993	0	1	0	1	0	0	0	0	8	0
corn	0.75	0.6	0.667	0.995	0	0	0	0	0	0	0	0	2	3

Tabela D.2.1: Performance por classe e Matriz de confusão para Palavras

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.955	0.955	0.955	0.967	191	5	0	0	3	0	1	0	0	0
earn	0.871	0.955	0.911	0.949	5	149	0	0	0	1	0	0	1	0
money-fx	0.789	0.6	0.682	0.966	1	2	15	1	0	2	3	1	0	0
grain	0.5	0.2	0.286	0.942	2	1	0	2	0	1	0	0	1	3
crude	0.828	0.828	0.828	0.95	0	3	0	0	24	0	0	2	0	0
trade	0.722	0.867	0.788	0.983	0	1	0	0	0	13	1	0	0	0
interest	0.762	0.571	0.653	0.911	0	6	4	0	1	1	16	0	0	0
ship	0.813	0.813	0.813	0.963	0	2	0	0	1	0	0	13	0	0
wheat	0.6	0.6	0.6	0.865	0	2	0	0	0	0	0	0	3	0
corn	0.25	0.333	0.286	0.747	1	0	0	1	0	0	0	0	0	1

Tabela D.2.2: Performance por classe e Matriz de confusão para Multi-palavras

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.886	0.97	0.926	0.944	194	3	1	0	0	2	0	0	0	0
earn	0.93	0.942	0.936	0.961	7	147	0	0	1	0	0	1	0	0
money-fx	0.652	0.6	0.625	0.916	4	2	15	0	0	1	3	0	0	0
grain	1	0.2	0.333	0.993	4	1	0	2	0	1	0	0	1	1
crude	0.905	0.655	0.76	0.919	1	2	2	0	19	0	0	5	0	0
trade	0.65	0.867	0.743	0.925	2	0	0	0	0	13	0	0	0	0
interest	0.842	0.571	0.681	0.89	2	2	5	0	1	2	16	0	0	0
ship	0.667	0.75	0.706	0.981	3	1	0	0	0	0	0	12	0	0
wheat	0.8	0.8	0.8	0.996	1	0	0	0	0	0	0	0	4	0
corn	0.5	0.333	0.4	0.934	1	0	0	0	0	1	0	0	0	1

Tabela D.2.3: Performance por classe e Matriz de confusão para sequências de 4-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.864	0.97	0.914	0.919	229	6	0	0	0	0	1	0	0	0
earn	0.937	0.862	0.898	0.942	18	119	0	0	1	0	0	0	0	0
money-fx	0.867	0.565	0.684	0.964	6	1	13	0	0	1	2	0	0	0
grain	0.75	0.429	0.545	0.96	2	0	0	3	0	1	0	0	1	0
crude	0.9	0.75	0.818	0.921	1	1	0	0	18	0	0	4	0	0
trade	0.733	0.846	0.786	0.966	2	0	0	0	0	11	0	0	0	0
interest	0.789	0.625	0.698	0.876	5	0	2	0	0	1	15	1	0	0
ship	0.545	0.6	0.571	0.98	2	0	0	0	1	0	1	6	0	0
wheat	0.875	0.875	0.875	0.999	0	0	0	1	0	0	0	0	7	0
corn	1	0.5	0.667	0.991	0	0	0	0	0	1	0	0	0	1

Tabela D.2.4: Performance por classe e Matriz de confusão para sequências de 5-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.805	0.958	0.875	0.901	206	6	0	0	0	1	1	0	1	0
earn	0.94	0.869	0.903	0.933	17	126	0	0	1	0	0	1	0	0
money-fx	0.882	0.714	0.789	0.954	3	0	15	0	0	2	1	0	0	0
grain	0.333	0.2	0.25	0.965	3	0	0	1	1	0	0	0	0	0
crude	0.667	0.4	0.5	0.878	8	0	0	0	10	0	0	7	0	0
trade	0.778	0.609	0.683	0.912	6	1	0	0	0	14	1	0	1	0
interest	0.813	0.565	0.667	0.897	7	0	2	0	0	1	13	0	0	0
ship	0.385	0.417	0.4	0.875	3	1	0	0	3	0	0	5	0	0
wheat	0.375	0.75	0.5	0.869	1	0	0	0	0	0	0	0	3	0
corn	0	0	0	0.813	2	0	0	2	0	0	0	0	3	0

Tabela D.2.5: Performance por classe e Matriz de confusão para sequências de 6-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.9	0.985	0.94	0.957	197	2	1	0	0	0	0	0	0	0
earn	0.937	0.949	0.943	0.966	6	148	0	0	2	0	0	0	0	0
money-fx	0.696	0.64	0.667	0.945	4	2	16	0	0	1	2	0	0	0
grain	1	0.1	0.182	0.995	4	1	0	1	2	1	0	0	0	1
crude	0.75	0.724	0.737	0.911	1	1	1	0	21	1	1	3	0	0
trade	0.765	0.867	0.813	0.939	1	1	0	0	0	13	0	0	0	0
interest	0.857	0.643	0.735	0.95	2	2	5	0	0	1	18	0	0	0
ship	0.786	0.688	0.733	0.989	2	1	0	0	2	0	0	11	0	0
wheat	1	0.8	0.889	0.998	1	0	0	0	0	0	0	0	4	0
corn	0.5	0.333	0.4	0.749	1	0	0	0	1	0	0	0	0	1

Tabela D.2.6: Performance por classe e Matriz de confusão para sequências de 4, 5 e 6-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.898	0.97	0.933	0.94	229	6	0	0	0	0	1	0	0	0
earn	0.938	0.884	0.91	0.946	15	122	0	0	0	1	0	0	0	0
money-fx	0.737	0.609	0.667	0.947	4	1	14	0	0	2	2	0	0	0
grain	0.75	0.429	0.545	0.992	0	0	0	3	0	1	1	0	1	1
crude	0.85	0.708	0.773	0.931	2	1	1	0	17	0	0	3	0	0
trade	0.688	0.846	0.759	0.943	1	0	0	0	0	11	1	0	0	0
interest	0.75	0.625	0.682	0.881	3	0	4	0	0	1	15	1	0	0
ship	0.6	0.6	0.6	0.987	1	0	0	0	3	0	0	6	0	0
wheat	0.889	1	0.941	0.999	0	0	0	0	0	0	0	0	8	0
corn	0.5	0.5	0.5	0.983	0	0	0	1	0	0	0	0	0	1

Tabela D.2.7: Performance por classe e Matriz de confusão para Pentagramas

D.3 Performance utilizando a técnica *Information Gain*

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.965	0.965	0.965	0.978	193	4	1	0	1	1	0	0	0	0
earn	0.902	0.961	0.931	0.964	3	148	1	0	1	1	0	0	0	0
money-fx	0.857	0.889	0.873	0.972	1	0	24	0	0	0	2	0	0	0
grain	0.8	0.571	0.667	0.984	0	2	0	4	0	0	0	0	1	0
crude	0.867	0.703	0.776	0.963	1	4	0	0	26	2	0	3	1	0
trade	0.789	0.833	0.811	0.942	1	2	0	0	0	15	0	0	0	0
interest	0.882	0.75	0.811	0.982	1	1	2	0	0	0	15	1	0	0
ship	0.667	0.727	0.696	0.917	0	1	0	0	2	0	0	8	0	0
wheat	0.636	0.7	0.667	0.993	0	2	0	1	0	0	0	0	7	0
corn	1	0.6	0.75	0.996	0	0	0	0	0	0	0	0	2	3

Tabela D.3.1: Performance por classe e Matriz de confusão para Palavras

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.969	0.945	0.957	0.973	189	8	0	0	1	1	1	0	0	0
earn	0.847	0.962	0.901	0.946	3	150	0	1	0	1	0	1	0	0
money-fx	0.583	0.56	0.571	0.947	0	2	14	1	0	2	5	1	0	0
grain	0	0	0	0.836	1	3	1	0	0	2	0	0	0	3
crude	0.889	0.828	0.857	0.952	0	3	1	0	24	0	0	1	0	0
trade	0.684	0.867	0.765	0.978	0	1	0	0	0	13	1	0	0	0
interest	0.682	0.536	0.6	0.889	1	4	7	0	1	0	15	0	0	0
ship	0.786	0.688	0.733	0.965	0	3	1	0	1	0	0	11	0	0
wheat	1	0.6	0.75	0.951	0	2	0	0	0	0	0	0	3	0
corn	0.25	0.333	0.286	0.569	1	1	0	0	0	0	0	0	0	1

Tabela D.3.2: Performance por classe e Matriz de confusão para Multi-palavras

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.886	0.97	0.926	0.944	194	3	1	0	0	2	0	0	0	0
earn	0.93	0.942	0.936	0.961	6	147	0	0	1	1	0	1	0	0
money-fx	0.682	0.6	0.638	0.906	4	2	15	0	0	1	3	0	0	0
grain	1	0.3	0.462	0.989	4	1	0	3	0	1	0	0	1	0
crude	0.955	0.724	0.824	0.937	1	2	1	0	21	1	0	3	0	0
trade	0.619	0.867	0.722	0.918	2	0	0	0	0	13	0	0	0	0
interest	0.85	0.607	0.708	0.907	2	2	5	0	0	2	17	0	0	0
ship	0.75	0.75	0.75	0.983	3	1	0	0	0	0	0	12	0	0
wheat	0.8	0.8	0.8	0.996	1	0	0	0	0	0	0	0	4	0
corn	1	0.333	0.5	0.925	2	0	0	0	0	0	0	0	0	1

Tabela D.3.3: Performance por classe e Matriz de confusão para sequências de 4-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.864	0.97	0.914	0.919	229	6	0	0	0	0	1	0	0	0
earn	0.937	0.862	0.898	0.942	18	119	0	0	1	0	0	0	0	0
money-fx	0.867	0.565	0.684	0.964	6	1	13	0	0	1	2	0	0	0
grain	0.75	0.429	0.545	0.959	2	0	0	3	0	1	0	0	1	0
crude	0.9	0.75	0.818	0.921	1	1	0	0	18	0	0	4	0	0
trade	0.733	0.846	0.786	0.966	2	0	0	0	0	11	0	0	0	0
interest	0.789	0.625	0.698	0.876	5	0	2	0	0	1	15	1	0	0
ship	0.545	0.6	0.571	0.98	2	0	0	0	1	0	1	6	0	0
wheat	0.875	0.875	0.875	0.999	0	0	0	1	0	0	0	0	7	0
corn	1	0.5	0.667	0.991	0	0	0	0	0	1	0	0	0	1

Tabela D.3.4: Performance por classe e Matriz de confusão para sequências de 5-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.809	0.963	0.879	0.904	207	5	0	0	0	1	1	0	1	0
earn	0.948	0.876	0.91	0.937	16	127	0	0	1	0	0	1	0	0
money-fx	0.882	0.714	0.789	0.955	3	0	15	0	0	2	1	0	0	0
grain	0.333	0.2	0.25	0.965	3	0	0	1	1	0	0	0	0	0
crude	0.667	0.4	0.5	0.879	8	0	0	0	10	0	0	7	0	0
trade	0.778	0.609	0.683	0.933	6	1	0	0	0	14	1	0	1	0
interest	0.813	0.565	0.667	0.875	7	0	2	0	0	1	13	0	0	0
ship	0.385	0.417	0.4	0.876	3	1	0	0	3	0	0	5	0	0
wheat	0.375	0.75	0.5	0.869	1	0	0	0	0	0	0	0	3	0
corn	0	0	0	0.813	2	0	0	2	0	0	0	0	3	0

Tabela D.3.5: Performance por classe e Matriz de confusão para sequências de 6-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.9	0.985	0.94	0.959	197	2	1	0	0	0	0	0	0	0
earn	0.932	0.962	0.946	0.97	5	150	0	0	1	0	0	0	0	0
money-fx	0.696	0.64	0.667	0.946	4	2	16	0	0	1	2	0	0	0
grain	1	0.1	0.182	0.995	4	1	0	1	2	1	0	0	0	1
crude	0.778	0.724	0.75	0.913	2	1	1	0	21	0	1	3	0	0
trade	0.813	0.867	0.839	0.94	1	1	0	0	0	13	0	0	0	0
interest	0.857	0.643	0.735	0.949	2	2	5	0	0	1	18	0	0	0
ship	0.786	0.688	0.733	0.989	2	1	0	0	2	0	0	11	0	0
wheat	1	0.6	0.75	0.997	1	1	0	0	0	0	0	0	3	0
corn	0.5	0.333	0.4	0.752	1	0	0	0	1	0	0	0	0	1

Tabela D.3.6: Performance por classe e Matriz de confusão para sequências de 4, 5 e 6-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.898	0.966	0.931	0.937	228	7	0	0	0	0	1	0	0	0
earn	0.938	0.884	0.91	0.946	15	122	0	0	0	1	0	0	0	0
money-fx	0.737	0.609	0.667	0.944	4	0	14	0	0	2	3	0	0	0
grain	0.75	0.429	0.545	0.992	0	0	0	3	0	1	1	0	1	1
crude	0.85	0.708	0.773	0.915	2	1	1	0	17	0	0	3	0	0
trade	0.688	0.846	0.759	0.943	1	0	0	0	0	11	1	0	0	0
interest	0.714	0.625	0.667	0.881	3	0	4	0	0	1	15	1	0	0
ship	0.6	0.6	0.6	0.986	1	0	0	0	3	0	0	6	0	0
wheat	0.889	1	0.941	0.999	0	0	0	0	0	0	0	0	8	0
corn	0.5	0.5	0.5	0.983	0	0	0	1	0	0	0	0	0	1

Tabela D.3.7: Performance por classe e Matriz de confusão para Pentagramas

Apêndice E

K-NN: Resultados por classe com a colecção R12

E.1 Performance obtida utilizando a técnica Terceiro Momento

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.742	0.771	0.756	0.889	118	12	9	1	7	0	3	3	0	0
earn	0.878	0.935	0.906	0.963	9	187	0	0	3	0	0	1	0	0
money-fx	0.387	0.444	0.414	0.809	9	0	12	0	3	1	2	0	0	0
grain	0.4	0.286	0.333	0.749	2	0	0	2	0	0	0	0	2	1
crude	0.419	0.361	0.388	0.859	12	3	2	0	13	1	1	3	1	0
trade	0.667	0.316	0.429	0.802	4	5	0	0	1	6	2	1	0	0
interest	0.526	0.476	0.5	0.832	2	2	5	0	1	1	10	0	0	0
ship	0.273	0.273	0.273	0.794	1	1	2	0	3	0	1	3	0	0
wheat	0.444	0.4	0.421	0.845	2	3	0	1	0	0	0	0	4	0
corn	0.5	0.2	0.286	0.829	0	0	1	1	0	0	0	0	2	1

Tabela E.1.1: Performance por classe e Matriz de confusão para Palavras

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	com
acq	0.633	0.741	0.683	0.852	100	24	3	0	1	2	1	1	3	0
earn	0.803	0.871	0.836	0.905	23	183	1	0	1	0	0	0	2	0
money-fx	0.37	0.313	0.339	0.747	9	5	10	0	2	0	5	0	1	0
grain	0.2	0.167	0.182	0.767	1	2	0	1	0	0	0	0	0	2
crude	0.706	0.387	0.5	0.881	5	3	2	2	12	2	1	3	0	1
trade	0.6	0.4	0.48	0.877	2	1	3	0	0	6	0	2	1	0
interest	0.5	0.32	0.39	0.795	7	5	4	1	0	0	8	0	0	0
ship	0.4	0.211	0.276	0.686	7	3	3	0	1	0	0	4	0	1
wheat	0.222	0.222	0.222	0.765	3	1	1	0	0	0	0	0	2	2
corn	0	0	0	0.806	1	1	0	1	0	0	1	0	0	0

Tabela E.1.2: Performance por classe e Matriz de confusão para Multi-palavras

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.71	0.705	0.707	0.839	110	34	2	0	1	2	4	2	1	0
earn	0.794	0.919	0.852	0.923	14	193	0	0	1	1	0	1	0	0
money-fx	0.692	0.563	0.621	0.845	4	4	18	0	0	0	4	2	0	0
grain	0.5	0.5	0.5	0.964	0	0	0	1	0	0	0	0	0	1
crude	0.833	0.37	0.513	0.829	10	1	0	0	10	1	2	3	0	0
trade	0.615	0.533	0.571	0.895	2	4	0	0	0	8	1	0	0	0
interest	0.4	0.333	0.364	0.796	6	4	5	0	0	0	8	1	0	0
ship	0.182	0.182	0.182	0.754	6	2	0	0	0	0	1	2	0	0
wheat	0.75	0.429	0.545	0.849	1	0	1	1	0	1	0	0	3	0
corn	0	0	0	0.823	2	1	0	0	0	0	0	0	0	0

Tabela E.1.3: Performance por classe e Matriz de confusão para sequências de 4-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.741	0.601	0.664	0.809	83	41	2	0	5	3	2	2	0	0
earn	0.766	0.903	0.829	0.865	16	213	2	0	3	1	1	0	0	0
money-fx	0.611	0.478	0.537	0.776	2	6	11	0	0	0	3	1	0	0
grain	1	0.5	0.667	0.854	0	1	0	4	0	1	0	0	2	0
crude	0.56	0.583	0.571	0.836	6	2	0	0	14	0	1	1	0	0
trade	0.6	0.692	0.643	0.905	1	1	0	0	0	9	1	1	0	0
interest	0.389	0.292	0.333	0.714	2	10	3	0	1	1	7	0	0	0
ship	0.167	0.111	0.133	0.739	1	4	0	0	2	0	0	1	1	0
wheat	0.556	0.625	0.588	0.907	0	0	0	0	0	0	3	0	5	0
corn	0	0	0	0.884	1	0	0	0	0	0	0	0	1	0

Tabela E.1.4: Performance por classe e Matriz de confusão para sequências de 5-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.696	0.584	0.635	0.796	80	41	6	0	4	2	2	1	1	0
earn	0.733	0.914	0.814	0.861	11	192	4	1	0	1	0	0	1	0
money-fx	0.308	0.276	0.291	0.737	3	7	8	0	1	1	7	2	0	0
grain	0.143	0.1	0.118	0.687	0	4	1	1	0	1	0	0	3	0
crude	0.381	0.308	0.34	0.682	8	2	1	2	8	1	0	2	2	0
trade	0.438	0.333	0.378	0.725	3	3	0	2	3	7	2	0	1	0
interest	0.214	0.125	0.158	0.661	4	8	5	0	3	1	3	0	0	0
ship	0.286	0.167	0.211	0.685	3	4	1	0	1	1	0	2	0	0
wheat	0.182	0.4	0.25	0.935	2	0	0	0	0	0	0	0	2	1
corn	0	0	0	0.796	1	1	0	1	1	1	0	0	1	0

Tabela E.1.5: Performance por classe e Matriz de confusão para sequências de 6-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.667	0.643	0.655	0.827	92	31	4	1	1	5	2	6	0	1
earn	0.781	0.858	0.818	0.891	24	182	0	0	1	1	0	4	0	0
money-fx	0.381	0.32	0.348	0.799	3	7	8	0	0	0	4	3	0	0
grain	0.167	0.143	0.154	0.745	1	1	0	1	0	0	0	1	1	2
crude	0.722	0.481	0.578	0.836	4	3	0	2	13	0	0	3	2	0
trade	0.579	0.478	0.524	0.753	4	3	1	1	0	11	0	1	1	1
interest	0.7	0.452	0.549	0.759	4	2	8	0	1	1	14	1	0	0
ship	0.05	0.1	0.067	0.553	5	2	0	0	1	1	0	1	0	0
wheat	0.333	0.4	0.364	0.786	1	1	0	0	1	0	0	0	2	0
corn	0.2	0.333	0.25	0.825	0	1	0	1	0	0	0	0	0	1

Tabela E.1.6: Performance por classe e Matriz de confusão para sequências de 4, 5 e 6-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.658	0.691	0.674	0.813	96	31	2	0	2	1	6	1	0	0
earn	0.726	0.855	0.785	0.834	20	183	1	1	1	1	4	2	1	0
money-fx	0.5	0.333	0.4	0.709	4	9	9	0	1	0	2	1	1	0
grain	0.5	0.167	0.25	0.594	2	1	0	1	0	2	0	0	0	0
crude	0.75	0.429	0.545	0.759	10	5	1	0	15	1	2	1	0	0
trade	0.417	0.238	0.303	0.727	5	7	1	0	0	5	2	0	1	0
interest	0.111	0.143	0.125	0.645	4	6	2	0	0	0	2	0	0	0
ship	0.286	0.105	0.154	0.628	3	9	2	0	0	1	0	2	2	0
wheat	0.375	0.429	0.4	0.798	2	0	0	0	1	0	0	0	3	1
corn	0	0	0	0.645	0	1	0	0	0	1	0	0	0	0

Tabela E.1.7: Performance por classe e Matriz de confusão para Pentagramas

E.2 Performance obtida utilizando a técnica *Chi Square*

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.842	0.985	0.908	0.982	197	2	1	0	0	0	0	0	0	0
earn	0.851	0.851	0.851	0.934	21	131	0	0	0	1	1	0	0	0
money-fx	0.655	0.704	0.679	0.878	2	1	19	0	0	0	5	0	0	0
grain	1	0.286	0.444	0.753	2	1	1	2	0	0	0	0	0	1
crude	0.909	0.541	0.678	0.921	6	5	2	0	20	1	0	3	0	0
trade	0.833	0.556	0.667	0.893	1	5	1	0	1	10	0	0	0	0
interest	0.588	0.5	0.541	0.871	2	5	3	0	0	0	10	0	0	0
ship	0.625	0.455	0.526	0.882	0	2	2	0	1	0	1	5	0	0
wheat	0.75	0.6	0.667	0.881	2	2	0	0	0	0	0	0	6	0
corn	0.667	0.4	0.5	0.736	1	0	0	0	0	0	0	0	2	2

Tabela E.2.1: Performance por classe e Matriz de confusão para Palavras

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.912	0.88	0.896	0.94	176	24	0	0	0	0	0	0	0	0
earn	0.668	0.853	0.749	0.876	10	133	3	1	3	0	3	1	0	2
money-fx	0.5	0.44	0.468	0.819	2	8	11	0	1	0	3	0	0	0
grain	0.333	0.1	0.154	0.599	2	5	0	1	0	1	0	0	0	1
crude	0.778	0.724	0.75	0.916	0	7	0	0	21	0	0	1	0	0
trade	0.714	0.333	0.455	0.915	2	4	2	0	1	5	1	0	0	0
interest	0.696	0.571	0.627	0.812	1	7	3	0	0	1	16	0	0	0
ship	0.6	0.188	0.286	0.717	0	8	3	1	1	0	0	3	0	0
wheat	1	0.8	0.889	0.854	0	1	0	0	0	0	0	0	4	0
corn	0.25	0.333	0.286	0.647	0	2	0	0	0	0	0	0	0	1

Tabela E.2.2: Performance por classe e Matriz de confusão para Multi-palavras

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.828	0.965	0.891	0.958	193	7	0	0	0	0	0	0	0	0
earn	0.855	0.872	0.863	0.935	15	136	1	0	1	1	2	0	0	0
money-fx	0.522	0.48	0.5	0.797	6	3	12	1	0	1	2	0	0	0
grain	0.833	0.5	0.625	0.835	2	1	0	5	1	1	0	0	0	0
crude	0.824	0.483	0.609	0.93	6	4	1	0	14	0	2	2	0	0
trade	0.75	0.6	0.667	0.936	2	3	0	0	0	9	0	1	0	0
interest	0.667	0.571	0.615	0.86	2	2	8	0	0	0	16	0	0	0
ship	0.727	0.5	0.593	0.879	5	0	1	0	1	0	1	8	0	0
wheat	0	0	0	0.693	1	2	0	0	0	0	1	0	0	1
corn	0.5	0.333	0.4	0.894	1	1	0	0	0	0	0	0	0	1

Tabela E.2.3: Performance por classe e Matriz de confusão para sequências de 4-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.814	0.966	0.884	0.921	228	5	0	0	0	1	2	0	0	0
earn	0.867	0.754	0.806	0.885	32	104	0	1	0	0	1	0	0	0
money-fx	0.733	0.478	0.579	0.828	5	4	11	0	0	1	2	0	0	0
grain	0.333	0.143	0.2	0.734	3	1	0	1	0	1	0	0	0	1
crude	0.882	0.625	0.732	0.936	5	4	0	0	15	0	0	0	0	0
trade	0.667	0.615	0.64	0.925	3	0	1	0	0	8	0	1	0	0
interest	0.773	0.708	0.739	0.937	2	1	3	0	0	1	17	0	0	0
ship	0.857	0.6	0.706	0.967	1	1	0	0	2	0	0	6	0	0
wheat	1	0.5	0.667	0.895	0	0	0	1	0	0	0	0	4	3
corn	0.2	0.5	0.286	0.877	1	0	0	0	0	0	0	0	0	1

Tabela E.2.4: Performance por classe e Matriz de confusão para sequências de 5-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.812	0.963	0.881	0.921	207	5	0	0	0	1	1	1	0	0
earn	0.829	0.8	0.814	0.891	27	116	1	0	0	0	1	0	0	0
money-fx	0.789	0.714	0.75	0.927	1	2	15	0	0	0	2	1	0	0
grain	0.4	0.4	0.4	0.767	1	2	0	2	0	0	0	0	0	0
crude	0.583	0.28	0.378	0.832	6	2	1	0	7	0	1	8	0	0
trade	0.909	0.435	0.588	0.885	6	4	0	0	1	10	2	0	0	0
interest	0.611	0.478	0.537	0.79	5	5	2	0	0	0	11	0	0	0
ship	0.286	0.333	0.308	0.874	2	1	0	1	4	0	0	4	0	0
wheat	0.667	0.5	0.571	0.808	0	0	0	0	0	0	0	0	2	2
corn	0.333	0.143	0.2	0.815	0	3	0	2	0	0	0	0	1	1

Tabela E.2.5: Performance por classe e Matriz de confusão para sequências de 6-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.851	0.945	0.896	0.948	189	10	0	0	1	0	0	0	0	0
earn	0.879	0.84	0.859	0.926	17	131	2	0	4	0	2	0	0	0
money-fx	0.652	0.6	0.625	0.883	1	3	15	0	0	0	6	0	0	0
grain	0.875	0.7	0.778	0.881	0	1	1	7	0	1	0	0	0	0
crude	0.63	0.586	0.607	0.916	5	4	0	0	17	0	0	3	0	0
trade	0.933	0.933	0.933	0.971	1	0	0	0	0	14	0	0	0	0
interest	0.68	0.607	0.642	0.885	6	0	4	0	1	0	17	0	0	0
ship	0.75	0.563	0.643	0.905	1	0	0	1	4	0	0	9	1	0
wheat	0.75	0.6	0.667	0.895	1	0	0	0	0	0	0	0	3	1
corn	0.5	0.333	0.4	0.895	1	0	1	0	0	0	0	0	0	1

Tabela E.2.6: Performance por classe e Matriz de confusão para sequências de 4, 5 e 6-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.822	0.958	0.885	0.912	226	8	0	0	1	0	1	0	0	0
earn	0.863	0.732	0.792	0.863	34	101	0	0	1	0	2	0	0	0
money-fx	0.708	0.739	0.723	0.955	3	2	17	0	0	1	0	0	0	0
grain	0.5	0.429	0.462	0.858	3	0	0	3	0	1	0	0	0	0
crude	0.8	0.667	0.727	0.92	4	1	0	0	16	0	1	2	0	0
trade	0.8	0.615	0.696	0.938	2	1	0	0	0	8	1	1	0	0
interest	0.737	0.583	0.651	0.847	2	1	7	0	0	0	14	0	0	0
ship	0.571	0.4	0.471	0.885	1	3	0	0	2	0	0	4	0	0
wheat	1	0.375	0.545	0.884	0	0	0	2	0	0	0	0	3	3
corn	0.25	0.5	0.333	0.994	0	0	0	1	0	0	0	0	0	1

Tabela E.2.7: Performance por classe e Matriz de confusão para Pentagramas

E.3 Performance obtida utilizando a técnica *Information Gain*

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.857	0.99	0.919	0.985	198	1	0	1	0	0	0	0	0	0
earn	0.889	0.831	0.859	0.93	23	128	0	0	1	1	1	0	0	0
money-fx	0.731	0.704	0.717	0.865	1	2	19	0	0	1	4	0	0	0
grain	0.571	0.571	0.571	0.69	2	0	0	4	0	0	0	0	0	1
crude	0.85	0.459	0.596	0.902	6	3	3	0	17	1	1	5	0	1
trade	0.727	0.444	0.552	0.828	1	2	1	0	1	8	2	3	0	0
interest	0.571	0.6	0.585	0.924	0	5	3	0	0	0	12	0	0	0
ship	0.5	0.727	0.593	0.918	0	1	0	0	1	0	1	8	0	0
wheat	0.833	0.5	0.625	0.865	0	2	0	2	0	0	0	0	5	1
corn	0.571	0.8	0.667	0.888	0	0	0	0	0	0	0	0	1	4

Tabela E.3.1: Performance por classe e Matriz de confusão para Palavras

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.878	0.9	0.889	0.93	180	15	2	0	1	0	1	0	1	0
earn	0.682	0.782	0.728	0.846	17	122	6	2	3	2	1	2	0	1
money-fx	0.25	0.28	0.264	0.674	1	10	7	1	1	0	5	0	0	0
grain	0	0	0	0.478	2	4	1	0	0	2	0	0	0	1
crude	0.7	0.483	0.571	0.848	1	8	1	0	14	3	1	0	1	0
trade	0.471	0.533	0.5	0.923	1	1	5	0	0	8	0	0	0	0
interest	0.6	0.429	0.5	0.751	2	7	3	0	0	2	12	2	0	0
ship	0.429	0.188	0.261	0.609	0	8	2	2	1	0	0	3	0	0
wheat	0.333	0.2	0.25	0.532	1	3	0	0	0	0	0	0	1	0
corn	0.333	0.333	0.333	0.718	0	1	1	0	0	0	0	0	0	1

Tabela E.3.2: Performance por classe e Matriz de confusão para Multi-palavras

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.82	0.98	0.893	0.955	196	4	0	0	0	0	0	0	0	0
earn	0.866	0.872	0.869	0.936	18	136	0	1	0	0	1	0	0	0
money-fx	0.409	0.36	0.383	0.771	5	4	9	1	0	1	5	0	0	0
grain	0.375	0.3	0.333	0.633	4	0	0	3	1	1	0	0	0	1
crude	0.889	0.552	0.681	0.935	6	5	1	0	16	0	0	1	0	0
trade	0.778	0.467	0.583	0.879	3	3	1	1	0	7	0	0	0	0
interest	0.714	0.536	0.612	0.85	1	1	11	0	0	0	15	0	0	0
ship	0.909	0.625	0.741	0.874	3	1	0	1	1	0	0	10	0	0
wheat	0	0	0	0.619	1	3	0	1	0	0	0	0	0	0
corn	0.5	0.333	0.4	0.884	2	0	0	0	0	0	0	0	0	1

Tabela E.3.3: Performance por classe e Matriz de confusão para sequências de 4-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.82	0.966	0.887	0.935	228	5	0	0	0	1	2	0	0	0
earn	0.861	0.761	0.808	0.902	30	105	0	0	0	1	2	0	0	0
money-fx	0.714	0.435	0.541	0.842	5	3	10	0	0	0	5	0	0	0
grain	0.5	0.286	0.364	0.815	0	1	0	2	0	1	2	0	0	1
crude	0.875	0.583	0.7	0.937	5	4	0	0	14	0	1	0	0	0
trade	0.727	0.615	0.667	0.892	3	0	1	1	0	8	0	0	0	0
interest	0.571	0.667	0.615	0.883	4	1	3	0	0	0	16	0	0	0
ship	1	0.4	0.571	0.907	2	2	0	0	2	0	0	4	0	0
wheat	1	0.75	0.857	0.931	0	1	0	0	0	0	0	0	6	1
corn	0	0	0	0.862	1	0	0	1	0	0	0	0	0	0

Tabela E.3.4: Performance por classe e Matriz de confusão para sequências de 5-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.766	0.977	0.859	0.918	210	3	0	0	0	1	0	1	0	0
earn	0.863	0.738	0.796	0.876	33	107	0	2	0	0	1	0	2	0
money-fx	0.867	0.619	0.722	0.901	2	3	13	0	0	0	2	1	0	0
grain	0.286	0.4	0.333	0.672	2	1	0	2	0	0	0	0	0	0
crude	0.583	0.28	0.378	0.813	7	1	0	0	7	0	1	8	1	0
trade	0.9	0.391	0.545	0.881	8	3	0	0	1	9	2	0	0	0
interest	0.6	0.391	0.474	0.769	7	5	2	0	0	0	9	0	0	0
ship	0.231	0.25	0.24	0.73	4	0	0	1	4	0	0	3	0	0
wheat	0.333	0.5	0.4	0.791	0	0	0	0	0	0	0	0	2	2
corn	0.5	0.286	0.364	0.826	1	1	0	2	0	0	0	0	1	2

Tabela E.3.5: Performance por classe e Matriz de confusão para sequências de 6-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.819	0.97	0.888	0.968	194	5	0	0	0	0	1	0	0	0
earn	0.899	0.801	0.847	0.942	20	125	1	0	2	0	8	0	0	0
money-fx	0.538	0.56	0.549	0.859	6	1	14	0	0	0	4	0	0	0
grain	0.8	0.4	0.533	0.723	3	1	1	4	0	1	0	0	0	0
crude	0.8	0.552	0.653	0.926	4	2	2	0	16	0	2	3	0	0
trade	0.769	0.667	0.714	0.925	2	2	0	0	0	10	1	0	0	0
interest	0.467	0.5	0.483	0.819	3	1	8	0	0	2	14	0	0	0
ship	0.75	0.563	0.643	0.816	4	0	0	1	2	0	0	9	0	0
wheat	1	0.6	0.75	0.843	0	1	0	0	0	0	0	0	3	1
corn	0.5	0.333	0.4	0.897	1	1	0	0	0	0	0	0	0	1

Tabela E.3.6: Performance por classe e Matriz de confusão para sequências de 4, 5 e 6-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.829	0.983	0.899	0.94	232	3	0	0	0	0	1	0	0	0
earn	0.931	0.783	0.85	0.905	27	108	0	0	1	0	2	0	0	0
money-fx	0.737	0.609	0.667	0.9	6	1	14	0	0	0	2	0	0	0
grain	0.667	0.571	0.615	0.851	0	1	0	4	0	1	0	0	1	0
crude	0.842	0.667	0.744	0.907	6	1	0	0	16	0	0	1	0	0
trade	0.889	0.615	0.727	0.914	3	1	0	0	0	8	1	0	0	0
interest	0.667	0.583	0.622	0.86	4	1	5	0	0	0	14	0	0	0
ship	0.833	0.5	0.625	0.84	2	0	0	0	2	0	0	5	1	0
wheat	0.667	0.5	0.571	0.833	0	0	0	1	0	0	1	0	4	2
corn	0.333	0.5	0.4	0.994	0	0	0	1	0	0	0	0	0	1

Tabela E.3.7: Performance por classe e Matriz de confusão para Pentagramas

Apêndice F

RIPPER: Resultados por classe para a colecção R12

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.9	0.945	0.922	0.952	189	10	0	0	0	0	1	0	0	0
earn	0.904	0.922	0.913	0.959	8	142	0	0	3	1	0	0	0	0
money-fx	0.875	0.778	0.824	0.941	2	1	21	0	0	1	2	0	0	0
grain	1	0.571	0.727	0.919	2	0	0	4	0	0	0	0	1	0
crude	0.846	0.595	0.698	0.915	1	1	1	0	22	1	3	8	0	0
trade	0.824	0.778	0.8	0.957	3	1	0	0	0	14	0	0	0	0
interest	0.667	0.6	0.632	0.874	5	1	2	0	0	0	12	0	0	0
ship	0.529	0.818	0.643	0.856	0	1	0	0	1	0	0	9	0	0
wheat	0.9	0.9	0.9	0.949	0	0	0	0	0	0	0	0	9	1
corn	0.833	1	0.909	0.999	0	0	0	0	0	0	0	0	0	5

Tabela F.1.1: Performance por classe e Matriz de confusão para Palavras

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.891	0.86	0.875	0.904	172	18	1	0	5	0	4	0	0	0
earn	0.798	0.859	0.827	0.91	9	134	4	0	5	1	2	1	0	0
money-fx	0.458	0.44	0.449	0.83	6	2	11	0	0	1	5	0	0	0
grain	0.5	0.1	0.167	0.638	0	3	0	1	1	1	0	0	1	3
crude	0.639	0.793	0.708	0.916	0	0	2	0	23	0	2	2	0	0
trade	0.667	0.667	0.667	0.858	2	2	0	0	0	10	0	1	0	0
interest	0.519	0.5	0.509	0.73	2	5	5	0	0	2	14	0	0	0
ship	0.667	0.5	0.571	0.837	2	2	1	1	2	0	0	8	0	0
wheat	0.8	0.8	0.8	0.858	0	1	0	0	0	0	0	0	4	0
corn	0.4	0.667	0.5	0.982	0	1	0	0	0	0	0	0	0	2

Tabela F.1.2: Performance por classe e Matriz de confusão para Multi-palavras

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.924	0.915	0.92	0.947	183	11	1	0	2	2	0	1	0	0
earn	0.854	0.897	0.875	0.939	9	140	0	1	1	2	3	0	0	0
money-fx	0.731	0.76	0.745	0.951	2	0	19	0	0	1	3	0	0	0
grain	0.333	0.1	0.154	0.759	2	5	0	1	0	1	0	0	1	0
crude	0.826	0.655	0.731	0.891	0	3	1	1	19	1	2	2	0	0
trade	0.632	0.8	0.706	0.919	0	2	0	0	1	12	0	0	0	0
interest	0.714	0.714	0.714	0.917	1	2	5	0	0	0	20	0	0	0
ship	0.833	0.938	0.882	0.993	0	1	0	0	0	0	0	15	0	0
wheat	0.833	1	0.909	0.999	0	0	0	0	0	0	0	0	5	0
corn	1	0.667	0.8	0.833	1	0	0	0	0	0	0	0	0	2

Tabela F.1.3: Performance por classe e Matriz de confusão para sequências de 4-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.887	0.928	0.907	0.923	219	12	2	0	0	1	2	0	0	0
earn	0.87	0.826	0.848	0.911	18	114	0	0	3	2	1	0	0	0
money-fx	0.75	0.783	0.766	0.976	1	0	18	0	1	1	2	0	0	0
grain	1	0.571	0.727	0.996	0	0	0	4	0	1	0	0	0	2
crude	0.727	0.667	0.696	0.936	1	1	0	0	16	3	0	3	0	0
trade	0.522	0.923	0.667	0.951	0	1	0	0	0	12	0	0	0	0
interest	0.615	0.333	0.432	0.771	7	2	4	0	0	3	8	0	0	0
ship	0.667	0.6	0.632	0.803	1	1	0	0	2	0	0	6	0	0
wheat	1	1	1	1	0	0	0	0	0	0	0	0	8	0
corn	0.5	1	0.667	0.999	0	0	0	0	0	0	0	0	0	2

Tabela F.1.4: Performance por classe e Matriz de confusão para sequências de 5-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.852	0.884	0.868	0.891	190	18	1	0	5	1	0	0	0	0
earn	0.855	0.897	0.875	0.918	11	130	0	0	1	0	2	0	1	0
money-fx	0.714	0.714	0.714	0.85	0	0	15	1	0	1	4	0	0	0
grain	0.25	0.2	0.222	0.751	2	0	0	1	1	0	0	0	1	0
crude	0.565	0.52	0.542	0.89	3	1	1	0	13	0	0	7	0	0
trade	0.9	0.783	0.837	0.921	3	1	1	0	0	18	0	0	0	0
interest	0.571	0.348	0.432	0.788	10	0	3	0	1	0	8	0	1	0
ship	0.588	0.833	0.69	0.907	1	0	0	0	1	0	0	10	0	0
wheat	0.4	0.5	0.444	0.931	1	0	0	1	0	0	0	0	2	0
corn	1	0.143	0.25	0.605	2	2	0	1	1	0	0	0	0	1

Tabela F.1.5: Performance por classe e Matriz de confusão para sequências de 6-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.929	0.91	0.919	0.956	182	13	1	0	2	0	1	1	0	0
earn	0.877	0.91	0.893	0.948	8	142	0	0	1	3	2	0	0	0
money-fx	0.739	0.68	0.708	0.951	2	0	17	0	0	1	5	0	0	0
grain	0.6	0.6	0.6	0.936	1	1	0	6	0	1	0	0	1	0
crude	0.864	0.655	0.745	0.882	1	2	1	2	19	1	0	3	0	0
trade	0.65	0.867	0.743	0.981	0	2	0	0	0	13	0	0	0	0
interest	0.714	0.714	0.714	0.926	1	2	4	0	0	1	20	0	0	0
ship	0.778	0.875	0.824	0.994	0	0	0	2	0	0	0	14	0	0
wheat	0.833	1	0.909	0.999	0	0	0	0	0	0	0	0	5	0
corn	1	0.667	0.8	0.833	1	0	0	0	0	0	0	0	0	2

Tabela F.1.6: Performance por classe e Matriz de confusão para sequências de 4, 5 e 6-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.894	0.928	0.911	0.927	219	14	2	0	1	0	0	0	0	0
earn	0.874	0.855	0.864	0.921	12	118	0	0	3	1	2	1	0	1
money-fx	0.727	0.696	0.711	0.879	3	1	16	0	0	1	2	0	0	0
grain	0	0	0	0.659	3	1	0	0	0	1	0	0	0	2
crude	0.783	0.75	0.766	0.905	1	1	0	0	18	2	0	2	0	0
trade	0.688	0.846	0.759	0.945	2	0	0	0	0	11	0	0	0	0
interest	0.789	0.625	0.698	0.883	5	0	4	0	0	0	15	0	0	0
ship	0.75	0.9	0.818	0.916	0	0	0	0	1	0	0	9	0	0
wheat	0.889	1	0.941	0.999	0	0	0	0	0	0	0	0	8	0
corn	0.25	0.5	0.333	0.989	0	0	0	0	0	0	0	0	1	1

Tabela F.1.7: Performance por classe e Matriz de confusão para Pentagramas

Apêndice G

SVM: Resultados por classe com a colecção R4

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.918	0.949	0.934	0.961	225	7	0	1	4	0	0	0	0	0
earn	0.971	0.973	0.972	0.97	7	503	4	0	1	1	0	1	0	0
money-fx	0.607	0.607	0.607	0.949	4	2	17	0	0	3	2	0	0	0
grain	0.667	0.286	0.4	0.778	1	2	0	2	0	0	1	1	0	0
crude	0.667	0.7	0.683	0.943	5	0	1	0	14	0	0	0	0	0
trade	0.875	0.875	0.875	0.991	0	3	1	0	1	35	0	0	0	0
interest	0.88	0.71	0.786	0.98	2	0	5	0	1	1	22	0	0	0
ship	0.714	0.714	0.714	0.89	1	1	0	0	0	0	0	5	0	0
wheat	0	0	0	?	0	0	0	0	0	0	0	0	0	0
corn	0	0	0	?	0	0	0	0	0	0	0	0	0	0

Tabela G.1.1: Performance por classe e Matriz de confusão para Palavras

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.899	0.938	0.918	0.948	652	28	2	0	5	3	0	5	0	0
earn	0.962	0.98	0.971	0.974	17	1055	1	0	2	0	0	1	0	0
money-fx	0.727	0.552	0.627	0.93	17	4	48	0	0	6	9	3	0	0
grain	1	0.6	0.75	0.83	2	1	0	6	0	0	1	0	0	0
crude	0.89	0.815	0.851	0.945	13	1	1	0	97	0	1	6	0	0
trade	0.797	0.787	0.792	0.98	8	3	2	0	1	59	1	1	0	0
interest	0.826	0.704	0.76	0.944	6	4	12	0	0	2	57	0	0	0
ship	0.515	0.472	0.493	0.924	10	1	0	0	4	4	0	17	0	0
wheat	0	0	0	?	0	0	0	0	0	0	0	0	0	0
corn	0	0	0	?	0	0	0	0	0	0	0	0	0	0

Tabela G.1.2: Performance por classe e Matriz de confusão para Multi-palavras

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.927	0.931	0.929	0.95	647	29	3	5	8	1	2	0	0	0
earn	0.942	0.985	0.963	0.965	15	1058	1	0	0	0	0	0	0	0
money-fx	0.784	0.667	0.72	0.961	10	10	58	0	1	3	4	1	0	0
grain	0.5	0.6	0.545	0.934	0	3	0	6	0	0	1	0	0	0
crude	0.823	0.782	0.802	0.965	9	9	1	1	93	2	0	4	0	0
trade	0.847	0.813	0.83	0.987	5	3	1	0	4	61	0	1	0	0
interest	0.879	0.63	0.734	0.918	7	9	10	0	0	4	51	0	0	0
ship	0.778	0.583	0.667	0.891	5	2	0	0	7	1	0	21	0	0
wheat	0	0	0	?	0	0	0	0	0	0	0	0	0	0
corn	0	0	0	?	0	0	0	0	0	0	0	0	0	0

Tabela G.1.3: Performance por classe e Matriz de confusão para sequências de 4-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.942	0.871	0.905	0.933	605	71	4	3	9	1	1	1	0	0
earn	0.892	0.984	0.936	0.939	13	1026	1	0	2	0	1	0	0	0
money-fx	0.714	0.575	0.637	0.921	5	14	50	0	2	7	8	1	0	0
grain	0.333	0.3	0.316	0.886	0	6	0	3	0	0	1	0	0	0
crude	0.842	0.807	0.824	0.962	9	6	1	2	96	2	0	3	0	0
trade	0.823	0.867	0.844	0.984	2	5	2	0	0	65	0	1	0	0
interest	0.814	0.593	0.686	0.886	2	14	12	1	0	3	48	1	0	0
ship	0.696	0.444	0.542	0.908	6	8	0	0	5	1	0	16	0	0
wheat	0	0	0	?	0	0	0	0	0	0	0	0	0	0
corn	0	0	0	?	0	0	0	0	0	0	0	0	0	0

Tabela G.1.4: Performance por classe e Matriz de confusão para sequências de 5-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.933	0.862	0.896	0.93	598	78	6	3	7	0	1	1	0	0
earn	0.876	0.982	0.926	0.93	14	980	1	0	2	0	1	0	0	0
money-fx	0.724	0.632	0.675	0.93	8	14	55	0	0	4	6	0	0	0
grain	0.182	0.2	0.19	0.823	1	4	0	2	0	3	0	0	0	0
crude	0.852	0.773	0.811	0.96	9	12	1	2	92	1	0	2	0	0
trade	0.797	0.787	0.792	0.979	5	5	1	3	2	59	0	0	0	0
interest	0.843	0.531	0.652	0.889	1	21	12	1	0	3	43	0	0	0
ship	0.85	0.472	0.607	0.823	5	5	0	0	5	4	0	17	0	0
wheat	0	0	0	?	0	0	0	0	0	0	0	0	0	0
corn	0	0	0	?	0	0	0	0	0	0	0	0	0	0

Tabela G.1.5: Performance por classe e Matriz de confusão para sequências de 6-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.874	0.861	0.868	0.908	596	77	5	6	6	0	1	1	0	0
earn	0.884	0.967	0.924	0.927	27	1035	4	0	2	2	0	0	0	0
money-fx	0.6	0.419	0.493	0.889	21	13	36	1	1	8	4	2	0	0
grain	0.385	0.5	0.435	0.785	0	3	0	5	0	0	1	1	0	0
crude	0.863	0.746	0.8	0.966	15	8	0	1	88	1	0	5	0	0
trade	0.721	0.653	0.685	0.897	8	14	3	0	0	49	0	1	0	0
interest	0.891	0.605	0.721	0.956	5	10	10	0	3	3	49	1	0	0
ship	0.353	0.167	0.226	0.675	10	11	2	0	2	5	0	6	0	0
wheat	0	0	0	?	0	0	0	0	0	0	0	0	0	0
corn	0	0	0	?	0	0	0	0	0	0	0	0	0	0

Tabela G.1.6: Performance por classe e Matriz de confusão para sequências de 4, 5 e 6-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.902	0.822	0.86	0.902	568	106	5	1	5	0	1	5	0	0
earn	0.839	0.971	0.901	0.904	25	987	3	0	0	0	1	0	0	0
money-fx	0.609	0.459	0.523	0.867	11	23	39	0	0	4	5	3	0	0
grain	0.6	0.3	0.4	0.795	1	3	0	3	0	0	1	2	0	0
crude	0.919	0.771	0.839	0.955	7	12	1	1	91	1	1	4	0	0
trade	0.862	0.667	0.752	0.948	6	15	2	0	0	50	1	1	0	0
interest	0.811	0.531	0.642	0.917	5	16	14	0	1	2	43	0	0	0
ship	0.444	0.333	0.381	0.816	7	14	0	0	2	1	0	12	0	0
wheat	0	0	0	?	0	0	0	0	0	0	0	0	0	0
corn	0	0	0	?	0	0	0	0	0	0	0	0	0	0

Tabela G.1.7: Performance por classe e Matriz de confusão para Pentagramas

Apêndice H

K-NN: Resultados por classe com a colecção R4

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.767	0.848	0.806	0.96	201	30	3	0	1	0	2	0	0	0
earn	0.875	0.938	0.906	0.97	26	485	4	0	0	0	1	1	0	0
money-fx	0.391	0.321	0.353	0.936	6	9	9	0	1	0	3	0	0	0
grain	0	0	0	0.751	2	3	0	0	0	0	1	1	0	0
crude	0.333	0.1	0.154	0.839	11	6	0	0	2	0	0	1	0	0
trade	1	0.375	0.545	0.898	7	12	1	0	2	15	1	2	0	0
interest	0.571	0.387	0.462	0.951	6	7	6	0	0	0	12	0	0	0
ship	0.167	0.143	0.154	0.897	3	2	0	0	0	0	1	1	0	0
wheat	0	0	0	?	0	0	0	0	0	0	0	0	0	0
corn	0	0	0	?	0	0	0	0	0	0	0	0	0	0

Tabela H.1.1: Performance por classe e Matriz de confusão para Palavras

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.637	0.948	0.762	0.937	659	27	4	0	0	5	0	0	0	0
earn	0.935	0.897	0.916	0.964	108	965	2	0	1	0	0	0	0	0
money-fx	0.407	0.128	0.195	0.749	67	2	11	1	0	3	2	0	0	0
grain	0	0	0	0.816	9	1	0	0	0	0	0	0	0	0
crude	0.652	0.126	0.211	0.842	76	23	1	0	15	4	0	0	0	0
trade	0.5	0.213	0.299	0.877	44	8	3	0	2	16	0	2	0	0
interest	0.926	0.309	0.463	0.832	41	4	3	0	4	4	25	0	0	0
ship	0	0	0	0.695	30	2	3	0	1	0	0	0	0	0
wheat	0	0	0	?	0	0	0	0	0	0	0	0	0	0
corn	0	0	0	?	0	0	0	0	0	0	0	0	0	0

Tabela H.1.2: Performance por classe e Matriz de confusão para Multi-palavras

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.816	0.663	0.732	0.942	461	225	3	0	2	0	4	0	0	0
earn	0.731	0.984	0.839	0.977	12	1054	4	0	0	0	1	0	0	0
money-fx	0.429	0.172	0.246	0.838	17	47	15	0	0	3	4	1	0	0
grain	1	0.4	0.571	0.826	1	4	0	4	0	0	1	0	0	0
crude	0.889	0.336	0.488	0.904	33	35	2	0	40	1	7	1	0	0
trade	0.75	0.24	0.364	0.871	15	37	2	0	0	18	1	2	0	0
interest	0.655	0.444	0.529	0.84	15	23	6	0	0	1	36	0	0	0
ship	0.2	0.028	0.049	0.8	11	16	3	0	3	1	1	1	0	0
wheat	0	0	0	?	0	0	0	0	0	0	0	0	0	0
corn	0	0	0	?	0	0	0	0	0	0	0	0	0	0

Tabela H.1.3: Performance por classe e Matriz de confusão para sequências de 4-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.906	0.444	0.596	0.92	307	375	5	0	1	1	2	0	0	0
earn	0.62	0.988	0.762	0.936	9	1010	1	0	0	1	1	0	0	0
money-fx	0.5	0.212	0.298	0.788	6	57	18	0	0	3	1	0	0	0
grain	1	0.2	0.333	0.804	1	7	0	2	0	0	0	0	0	0
crude	0.961	0.412	0.576	0.906	6	58	2	0	49	1	2	1	0	0
trade	0.65	0.173	0.274	0.836	4	54	3	0	0	13	1	0	0	0
interest	0.811	0.37	0.508	0.843	4	39	7	0	0	1	30	0	0	0
ship	0.75	0.083	0.15	0.774	2	30	0	0	1	0	0	3	0	0
wheat	0	0	0	?	0	0	0	0	0	0	0	0	0	0
corn	0	0	0	?	0	0	0	0	0	0	0	0	0	0

Tabela H.1.4: Performance por classe e Matriz de confusão para sequências de 5-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.865	0.334	0.482	0.884	230	450	4	0	1	1	2	1	0	0
earn	0.581	0.991	0.732	0.91	5	954	2	0	0	0	1	1	0	0
money-fx	0.548	0.2	0.293	0.786	9	52	17	0	1	1	3	2	0	0
grain	0.8	0.4	0.533	0.855	0	4	0	4	0	1	1	0	0	0
crude	0.925	0.311	0.465	0.889	11	61	0	1	37	1	6	2	0	0
trade	0.708	0.227	0.343	0.85	6	49	1	0	0	17	0	2	0	0
interest	0.675	0.333	0.446	0.804	5	40	6	0	0	3	27	0	0	0
ship	0.2	0.056	0.087	0.748	0	32	1	0	1	0	0	2	0	0
wheat	0	0	0	?	0	0	0	0	0	0	0	0	0	0
corn	0	0	0	?	0	0	0	0	0	0	0	0	0	0

Tabela H.1.5: Performance por classe e Matriz de confusão para sequências de 6-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.765	0.775	0.77	0.922	527	118	18	0	8	1	1	7	0	0
earn	0.804	0.933	0.864	0.952	55	934	4	0	2	2	4	0	0	0
money-fx	0.38	0.366	0.373	0.858	13	20	30	0	4	5	9	1	0	0
grain	1	0.222	0.364	0.887	3	1	0	2	1	0	2	0	0	0
crude	0.569	0.25	0.347	0.852	37	38	6	0	29	0	2	4	0	0
trade	0.524	0.147	0.229	0.816	27	29	5	0	2	11	1	0	0	0
interest	0.667	0.5	0.571	0.919	14	11	12	0	1	1	40	1	0	0
ship	0.188	0.083	0.115	0.672	13	10	4	0	4	1	1	3	0	0
wheat	0	0	0	?	0	0	0	0	0	0	0	0	0	0
corn	0	0	0	?	0	0	0	0	0	0	0	0	0	0

Tabela H.1.6: Performance por classe e Matriz de confusão para sequências de 4, 5 e 6-caracteres

					Classificado como									
Classe	**Precisão**	***Recall***	**F- Measure**	**Área ROC**	acq	earn	money-fx	grain	crude	trade	interest	ship	wheat	corn
acq	0.836	0.658	0.736	0.902	432	197	10	0	10	1	3	4	0	0
earn	0.692	0.96	0.804	0.917	24	815	4	0	3	1	1	1	0	0
money-fx	0.482	0.346	0.403	0.817	13	25	27	0	6	3	1	3	0	0
grain	1	0.2	0.333	0.875	3	4	0	2	0	0	1	0	0	0
crude	0.681	0.419	0.519	0.88	18	38	5	0	49	2	1	4	0	0
trade	0.538	0.187	0.277	0.787	11	45	3	0	1	14	1	0	0	0
interest	0.784	0.372	0.504	0.803	7	31	7	0	1	3	29	0	0	0
ship	0	0	0	0.759	9	23	0	0	2	2	0	0	0	0
wheat	0	0	0	?	0	0	0	0	0	0	0	0	0	0
corn	0	0	0	?	0	0	0	0	0	0	0	0	0	0

Tabela H.1.7: Performance por classe e Matriz de confusão para Pentagramas

Bibliografia

Abreu, Marjory Cristiany Da Costa. "Analisando o desempenho do ClassAge: Um Sistema Multiagentes para Classificação de Padrõesl." Universidade Federal do Rio Grande do Norte, 2006.

Aires, J., G. P. Lopes, and J. F. Silva. "Efficient multi-word expressions using suffix arrays and related structures." Procceding of the 2nd ACM workshop on Improving non english websearching, pp.1-8,2008.

Apté, Chidanand, Fred Damerau, and Sholom M. Weiss. "Automated learning of decision rules for text categorization." ACM Transactions on Information Systems, Vol. 12, pp. 233 - 251, 1994.

Baranoski, Francis L. "Verificação da Autoria em Documentos Manuscritos Usando SVM." Dissertação apresentada ao Programa de Pós-Graduação em Informática Aplicada de Pontifícia, Universidade Católica do Panamá, 2005.

Basili, Roberto, Marco Cammisa, and Alessandro Mochitti. "A semantic kernel to exploit linguistic knowledge." Italian Association for Artificial Intelligence. Congress No9, Milan, pp. 290-302, Setembro, 2005.

Batista, Gustavo, and Maria Carolina Monard. "A Study of K-Nearest Neighbour as an Imputation Method." Soft Computing Systems: Design, Management and Applications, pp.251-260,2003.

"Biblioteca Digital do Brasil." http://www.bn.br/bndigital/, 2008.

Bikel, Daniel M., Richard Schwartz, and Ralph M. Weischedel. "An Algorithm that Learns What's in a Name." Machine Learning, nº34, pp. 211-231, 1999.

Blum, Avrim L., and Pat Langley. "Selection of relevant features and examples in machine learning." Artificial Intelligence, No. 1, pp. 245-271, Dezembro, 1997.

Braga, Ana Cristina Silva. "Curvas ROC: Aspectos Funcionais e Aplicações." Universidade do Minho, Braga, 2000.

"British National Corpus." (http://www.wlv.ac.uk/~in8113/data/bnc.tar.gz, Dezembro, 2008).

Broglio, John, James P. Callan, and W. Bruce Croft. "Inquery system overview." Proceedings of the TIPSTER Text Program, pp. 47-67, 1994.

Carletta, Jean. "Assessing agreement on classification tasks: the kappa statistic." Computational Linguistis, Vol.22, pp. 249-254, 1996.

Cohen, Jacob. "A coefficient of agreement for nominal scales." Educational and Psychological Measurement, Vol.20, pp.37-46, 1960.

Cohen, W. William, and Yoram Singer. "Context-Sensitive Learning Methods for Text Categorization." ACM Transactions on Information Systems, nº17, April, 1999.

"Datasets from Some Distributional Similarity Experiments." http://www.cs.cornell.edu/home/llee/data/sim.html, Dezembro, 2008.

Debole, Franca, and Fabrizio Sebastiani. "Supervised Term Weighting for Automated Text Categorization." Proceedings of the 2003 ACM symposium on Applied computing, pp. 784 - 788, 2003.

Fonseca, José Manuel Matos Ribeiro da. "Indução de Árvores de Decisão." Universidade Nova de Lisboa Faculdade de Ciências e Tecnologia, Departamento de Informática, 1994.

Galavotti, Luigi, Fabrizio Sebastiani, and Maria Simi. "Experiments on the Use of Feature Selection and Negative Evidence in Automated Text Categorization." Proceedings of ECDL-00, 4th European Conference on Research and Advanced Technology for Digital Libraries, pp. 59-68, 2000.

Galho, Thaís Silva, and Silvia Maria Wanderley Moraes. "Categorização Automática de Texto Utilizando Lógica Difusa." Universidade Luterana do Brasil (ULBRA), 2003.

Gonçalves, Teresa, and Paulo Quaresma. "Evaluating preprocessing techniques in a text classification problem." XXV Congresso da Sociedade Brasileira de Computação, pp. 841-850, Julho, 2005.

Hall, Mark A., and Geoffrey Holmes. "Benchmarking attribute selection techniques for discrete class data mining." IEEE Transactions On Knowledge And Data Engineering, nº15, pp. 1437-1447, 2003.

Hashimoto, K., and T. Yukawa. "Term weighting classification system using the chi-square statistic for the classification subtask at ntcir-6 patent retrieval task." Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access (NTCIR'07), pp. 385-389, 2007.

How, Bong Chih, and K. Narayanan. "An Empirical Study of Feature Selection for Text Categorization based on Term Weightage." Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 599-602, 2004.

Huisman, M. "Imputation of Missing Item Responses: Some Simple Techniques." Quality and Quantity, nº 4, Vol. 34, pp. 331-351, November, 2004.

"Internet Archive." http://www.archive.org/stream/teacherswordbook00thoruoft/teacherswordbook00thoruoft_djvu.txt, 2009.

Jacquemin, Christian, Judith L. Klavans, and Evelyne Tzoukermann. "Expansion of multi-word terms for indexing and retrieval using morphology and syntax." In proceedings of the 35th Annual Meeting of the ACL, pp. 24-31, 1997.

Käding, J. "Häufigkeitswörterbuch der deutschen Sprache." Steglitz, 1897.

Krkoska, Michael, Viktor Pekar, and Steffen Staab. "Feature Weighting for Co-occurrence-based Classification of Words." International Conference On Computational Linguistics, No. 799, 2004.

Krovetz, Robert. "Viewing morphology as an inference process." Department of Computer Science, University of Massachusetts, Amherst, 1993.

Kumaran, Giridhar, and James Allan. "Text classification and named entities for new event detection." Annual ACM Conference on Research and Development in Information Retrieval, pp. 297-304, 2004.

Lavesson, Niklas, and Paul Davidsson. "Quantifying the Impact of Learning Algorithm Parameter Tunning." Mälardelen Univerisity, Blekinge Institute of Technology, 2005.

Lewis, David D. "Representation and learning in information retrieval, PhD thesis." Department of Computer Science, University of Massachusetts, 1992.

Marques, Nuno, and José Gabriel Lopes. "Tagging With Small Training Corpora." Proceedings of the International Conference on Intelligent Data Analysis (IDA'01), pp. 63-72, 2001.

Mathieu, Benoit, Romanic Besancon, and Christian Fluhr. "Multilingual document clusters discovery." RIAO, 2004.

Mladenic, Dunja. "Feature subset selection in text-learning." Proceedings of the 10th European Conference on Machine Learning, pp. 95-100, 1998.

Muscat, Robert. "Automatic Document Clustering Using Topic Analysis." Master's thesis, University of Malta, 2005.

Myrtveit, Ingunn, Erik Stensrud, and Ulf H. Olsson. "Analyzing Data Sets with Missing Data: An Empirical Evaluation of Imputation Methods and Likelihood-Based Methods." IEEE Transactions on Software Engineering, Nº 11, Vol. 27, pp. 999 - 1013, 2001.

Nallapati, Ramesh. "Semantic language models for Topic Detection and Tracking." Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language, 2003.

Oren, Zamir, and Etzioni Oren. "Web Document Clustering: A Feasibility Demonstration." Department of Computer Science and Engineering, University of Washington, pp. 46-54, 1998.

Ozgur, Arzucan. "Supervised and Unsupervised Machine Learning Techniques for Text Document Categorization." B.S. in Computer Engineering, Bogazi¸ci University, 2004.

Papka, Ron, and James Allan. "Document classification using multiword features." Proceedings of the seventh international conference on Information and knowledge management, pp. 124-131, 1998.

Pons-Porrata, Aurora, Rafael Berlanga-Llavori, and José Ruiz-Shulcloper. "Topic discovery based on text mining techniques." Information Processing and Management: an International Journal, Vol. 43, pp. 752-768, 2007.

Pop, Ian. "An approach of the Naive Bayes classifier for the document classification." General Mathematics No.4, nº14, pp. 135-138, 2007.

Rish, Irina, Joseph Hellerstein, and Jayram Thathachar. "An analysis of data characteristics that affect Naïve Bayes performance." IBM T.J. Watson Research Center, 2001.

Sambasivam, Samuel, and Nick Theodosopoulos. "Advanced Data Clustering Methods of Mining Web Documents." Issues in Informing Science & Information Technology, nº3, pp.563-579, 2006.

Sardinha, Tony Berber. "Linguística de Corpus Histórico e Problemática." D.E.L.T.A., nº16, pp. 323-367, 2000.

Schapire, Robert E., and Yoram Singer. "BoosTexter: a boosting-based system for text categorization." Machine Learning, nº 39, pp. 135-168, Maio, 2000.

Sebastiani, Fabrizio. "A Tutorial on Automated Text Categorization." Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence, pp. 7-35, 1999.

Sebastiani, Fabrizio. "Machine Learning in Automated Text Categorization." ACM Computing Surveys , No. 1, 34, pp. 1-47, 2002.

Sebastiani, Fabrizio, and Franca Debole. "An analysis of the relative hardness of Reuters-21578 subsets." Journal of the American Society for Information Science and Technology, pp.584-596, 2005.

Shepperd, M. J., and M. H. Cartwright. "Dealing with Missing Software Project Data." Proceedings of the 9th International Symposium on Software Metrics, pp.154, 2003.

Silva, Joaquim Ferreira, Gaël Dias, Sylvie Guilloré, and José Gabriel Pereira Lopes. "Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units." Proceedings of the 9th Portuguese Conference on Artificial Intelligence: Progress in Artificial Intelligence, pp. 849-849, 1999.

Silva, Joaquim, João Mexia, Carlos A. Coelho, and Gabriel Lopes. "Multilingual Document Clustering and Data Transformations." Progress in Artificial Intelligence, Springer-Verlag, 2258 LNAI, pp. 74-87, 2001.

Smith, Lindsay I. "A tutorial on Principal Components Analysis." http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf, 7 Dezembro 2008.

Soares, Jorge Abreu. "Pré-Processamento em Mineração de Dados: Um Estudo Comparativo em Complementação." Universidade Federal do Rio de Janeiro, COPPE, 2007.

Sparck, K. Jones, and S. E. Robertson. "Relevance weighting of search terms." Journal of the American Society for Information Science, vol. 27,pp. 129-46, 1976.

Strijbos, Jan-Willem, Rob L. Martens, Frans J. Prins, and Wim M.G. Jochems. "Content analysis: what are they talking about?" Methodological issue in researching CSCL, pp.29-48, 2006.

"SVMLight - Support Vector Machine." http://www.cs.cornell.edu/People/tj/svm_light/, 2008.

Teevan, Jaime, and Rennie D. M. Jason. "Tackling the Poor Assumptions of Naive Bayes Text Classifiers." In Proceedings of the Twentieth International Conference on Machine Learning, pp. 616-623, 2003.

"The European Library." http://search.theeuropeanlibrary.org/portal/en/index.html, 2008.

Thorsten, Joachims. "Text Categorization with Support Vector Machines: Learning with Many Relevant Features." University of Dortmund, Computer Cience Department, 1998.

Valente, Iolanda Sofia Rendeiro. "Comparação entre Formatos de Classificação CDD, CDU e LCC." Junho, 2003.

"Wikipedia." http://en.wikipedia.org/wiki/Lemma_(linguistics), 2008.

"Wikipedia." http://en.wikipedia.org/wiki/Word_stem, 2008.

"Wikipedia." http://pt.wikipedia.org/wiki/Música, 2008.

"Wikipedia." http://en.wikipedia.org/wiki/Lazy_learning, 2009.

"Wikipedia." http://en.wikipedia.org/wiki/Part-of-speech_tagging, 2009.

"Wikipedia." http://en.wikipedia.org/wiki/Part-of-speech_tagging, 2008.

"Wikipedia." http://pt.wikipedia.org/wiki/Token, 2009.

"Wikipedia." http://en.wikipedia.org/wiki/KNN, 2009.

"Wikipedia." http://pt.wikipedia.org/wiki/Open_source, 2008.

"Wikipedia." http://pt.wikipedia.org/wiki/Economia, 2008.

"Wikipedia." http://en.wikipedia.org/wiki/Outliers, 2008.

Witten, Ian H., Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, and Sally Jo Cunningham. "Weka: Pratical Machine Learning Tools and Techniques with Java Implementation." Proc ICONIP/ANZIIS/ANNES99 Future Directions for Intelligent Systems and Information Sciences, pp.192-196, 1999.

Wong, Wai-chiu, and Ada Wai-chee Fu. "Incremental Document Clustering for Web Page Classification." Department of Computer Science and Engineering, 2000.

"WordNet - a lexical database for the English language." http://wordnet.princeton.edu/, Dezembro, 2008.

Xu, J., J. Broglio, and B. Croft. "The Design and Implementation of a Part of Speech Tagger for English." University of Massachusetts, 1994.

Yiming, Yang, and O. Pedersen Jan. "A Comparative Study on Feature Selection in Text Categorization." Proceedings of the Fourteenth International Conference on Machine Learning, pp. 412-420, 1997.

Zelikovitz, Sarah, and Haym Hirsh. "Using LSI for text classification in the presence of background text." Proceedings of CIKM-01, 10th ACM International Conference on Information and Knowledge Management, pp. 113-118, 2001.

Zheng, Zhaohui, and Rohini Srihari. "Optimally Combining Positive and Negative Features for Text Categorization." In Workshop for Learning from Imbalanced Datasets II, Proceedings of the ICML, 2003.

Zheng, Zhaohui, Rohini Srihari, and Sargur Srihari. "A Feature Selection Framework for Text Filtering." Proceedings of the Third IEEE International Conference on Data Mining, 2003.